Analyse et indexation de textes scientifiques

Florian Boudin

HABILITATION A DIRIGER DES RECHERCHES

NANTES UNIVERSITE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Florian Boudin

Analyse et indexation de textes scientifiques

Thèse présentée et soutenue à Nantes, le 20 juin 2023
Unité de recherche : LS2N (UMR 6004)

Composition du Jury :

Présidente et rapportrice :	Aurélie Névéol	Directrice de recherche CNRS, LISN Paris-Saclay
Rapporteurs :	Antoine Doucet	Professeur des universités, La Rochelle Université
	Jacques Savoy	Professeur titulaire, Université de Neuchâtel
Examinateurs :	Béatrice Daille	Professeure des universités, Nantes Université
	Richard Dufour	Professeur des universités, Nantes Université

Avant-propos

Ce manuscrit présente les travaux de recherche que j'ai réalisés au sein des différents laboratoires et équipes de recherche auxquels j'ai été rattaché depuis ma soutenance de thèse en 2008 (RALI de l'Université de Montréal de 2009 à 2010, LIA de l'Université d'Avignon de 2010 à 2011, puis équipe TALN au sein du LS2N de Nantes Université depuis 2011 ponctué d'un séjour de recherche au Aizawa-lab du National Institute of Informatics à Tokyo en 2019). Le « *nous* » y est utilisé car ces travaux sont évidemment le fruit de nombreuses collaborations locales et extérieures avec des chercheur.e.s et des étudiant.e.s. Je tiens ici à remercier tous ceux avec qui j'ai eu la chance de travailler, et en particulier les doctorants que j'ai co-encadré : Adrien Bougouin, Ygor Gallina, Maël Houbre et Léane Jourdan. J'adresse évidemment mes remerciements aux membres de l'équipe TALN, qui m'ont offert un environnement de recherche dynamique et épanouissant. J'adresse en particulier mes remerciements à Béatrice Daille, qui m'a poussé, soutenu et accompagné dans ce projet d'habilitation à diriger des recherches.

Ce manuscrit ne se veut pas exhaustif dans la portée ni dans la précision des travaux qui y sont présentés, mais propose un aperçu des activités de recherche que nous avons menées et de leur évolution au fil des années. Il témoigne des allers-retours constants que nous avons effectués entre les thématiques du Traitement Automatique des Langues et de la Recherche d'Information, et du positionnement singulier de nos travaux, au croisement de ces deux thématiques.

Table des Matières

I am looking for someone to share in an adventure that I am arranging, and it's very difficult to find anyone.

J.R.R. Tolkien, The Hobbit

1 Introduction

Sommaire

Les *bibliothèques numériques* occupent une place fondamentale dans l'organisation, la conservation et la mise à disposition des documents numériques. Avec l'accroissement rapide et continu du nombre de documents disponibles, la question de l'indexation, et plus généralement de la recherche de documents, revêt une dimension toute particulière. Cette question se pose avec d'autant plus d'acuité dans le monde scientifique où les bibliothèques numériques (e. g. les archives ouvertes `arXiv` et HAL, l'ACM Digital Library, ou PubMed), qui constituent aujourd'hui le point d'entrée principal au savoir scientifique, voient leur taille augmenter de façon considérable (Bornmann and Mutz, 2015). Ainsi, les activités essentielles à la recherche scientifique que sont la recherche bibliographique ou la veille scientifique demandent une quantité de travail de plus en plus importante. Simplifier et faciliter l'accès aux articles scientifiques est plus que jamais un enjeu majeur pour la communauté scientifique, et fait naturellement l'objet d'une attention soutenue auprès des chercheurs et des industriels du secteur académique.

Ce manuscrit synthétise les travaux que nous avons menés pour répondre à cet enjeu, avec pour fil conducteur la problématique de l'indexation automatique de documents par mots-clés. Les travaux qui y sont présentés se situent à la croisée de deux thématiques de recherche : celle du Traitement Automatique des Langues (TAL) qui concerne l'analyse, la compréhension et la production de langage naturel, et celle de la Recherche d'Information (RI) qui étudie la manière de retrouver des informations dans une collection de documents. Bien que ces deux thématiques entretiennent des rapports étroits, les communautés scientifiques qui s'y rapportent ont longtemps évolué séparément. Il en résulte naturellement des codes et des pratiques différentes que nous nous efforçons d'harmoniser au travers de ce manuscrit. Cette introduction précise le contexte dans lequel nos travaux s'inscrivent et donne les clés de lecture nécessaires à leur appréhension. Pour cela, nous faisons une brève revue de la littérature

autour des questions de la recherche bibliographique (§1.1), de l'indexation automatique par mots-clés (§1.2), et des méthodes de graphes pour l'ordonnancement de texte (§1.3). Nous terminons ce chapitre par l'organisation détaillée du manuscrit (§1.4) et par la projection chronologique de notre production scientifique sur les différents chapitres (Figure 1.4).

1.1. Recherche bibliographique

La recherche bibliographique consiste à identifier les documents dans la littérature scientifique (e. g. articles, ouvrages, thèses) en rapport avec un sujet d'étude ou plus largement un domaine de recherche. Cette tâche, auparavant bornée au catalogue des bibliothèques physiques, s'étend maintenant à celui massif et grandissant des bibliothèques numériques. Il s'agit en pratique d'une tâche de recherche d'information sur de grandes collections de documents composées essentiellement de notices bibliographiques (i. e. fiches descriptives regroupant les informations telles que les auteurs, le titre, le résumé). L'indexation des notices bibliographiques en lieu et place des textes intégraux s'explique avant tout par des raisons de licence (droits d'accès aux textes) et de limitations de ressources (en stockage, en calcul) (Huang et al., 2019). Il est néanmoins intéressant de noter que des travaux montrent que la précision des systèmes de recherche d'information est meilleure sur les notices que sur les textes intégraux (Lin, 2009).

Malgré l'émergence de moteurs de recherche dédiés couramment adossés à des systèmes de recommandation (e.g. Google Scholar, Microsoft Academic, Semantic Scholar), parcourir la littérature scientifique reste une activité laborieuse et chronophage (Gusenbauer and Haddaway, 2020). La première raison tient à l'explosion de la production scientifique mondiale qui noie les chercheurs sous

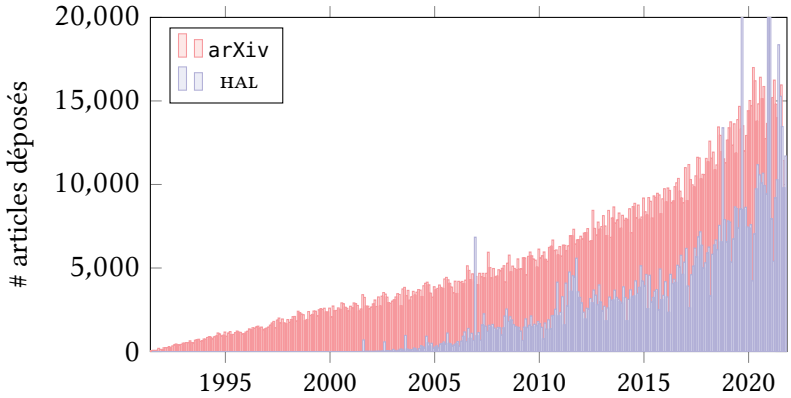


Figure 1.1. – Nombre d'articles scientifiques déposés chaque mois dans les archives ouvertes arXiv (données tirées de https://arxiv.org/stats/monthly_submissions) et HAL (données calculées avec https://gist.github.com/boudinfl/31a2bac37192960cba595c5c8f8e0166).

le nombre de résultats retournés. L'envolée du nombre d'articles déposés dans les archives ouvertes `arXiv` et HAL présentée dans la Figure 1.1 illustre parfaitement ce phénomène. Dans le seul domaine biomédical, plus d'un million d'articles sont déposés chaque année dans PubMed, soit en moyenne deux articles par minute (Landhuis, 2016). Ainsi, les moteurs de recherche retournent couramment plusieurs milliers de résultats pour une requête qu'il faut ensuite trier et filtrer (e.g. la requête "*named entity recognition*" retourne plus de 74 000 articles dans Google Scholar). L'autre raison tient à l'indexation souvent fragmentaire (i. e. qui s'appuie seulement sur la notice bibliographique) et insuffisamment interdisciplinaire (i. e. qui s'appuie sur des termes spécifiques au domaine) des textes scientifiques qui entraîne des problèmes de silence, se traduisant par résultats pertinents non retournés (Snyder, 2019).

Pour remédier à ces écueils, une solution consiste à enrichir les métadonnées des documents scientifiques pour améliorer leur accessibilité et leur diffusion. Pour cela, différents outils issus du TAL peuvent être utilisés comme des systèmes de catégorisation de textes (Howard and Ruder, 2018; Uban et al., 2021), de détection des références bibliographiques (Councill et al., 2008) ou d'extraction d'information (Peng and McCallum, 2004; Singh et al., 2016; Hou et al., 2019). Nos travaux s'inscrivent dans cette veine et concernent le marquage automatique des unités textuelles importantes dans les textes scientifiques, et plus particulièrement la génération automatique de mots-clés.

1.2. Indexation automatique par mots-clés

Les *mots-clés*, également appelés termes-clés ou descripteurs dans la littérature, sont des mots ou expressions polylexicales qui décrivent les principaux sujets abordés dans un document. Comme cela est illustré par l'exemple de la Figure 1.2, les mots-clés donnent une vue synthétique et condensée du contenu d'un document. Ils permettent de ce fait d'enrichir l'indexation des documents dans les bibliothèques numériques et, par ricochet, d'accroître l'efficacité des moteurs de recherche (Fagan, 1987; Zhai, 1997; Gutwin et al., 1999; Jones and Staveley, 1999). Les mots-clés sont également utilisés

Inverse problems for a mathematical model of ion exchange in a compressible ion exchanger

S. R. Tuikina
Computational Mathematics and Modeling, volume 13, pages 159–168 (2002)

A mathematical model of ion exchange is considered, allowing for ion exchanger compression in the process of ion exchange. Two inverse problems are investigated for this model, unique solvability is proved, and numerical solution methods are proposed. The efficiency of the proposed methods is demonstrated by a numerical experiment.

mots-clés : computability – inverse problems – ion exchange – mathematical programming

Figure 1.2. – Exemple de document (notice bibliographique) de la collection de test Inspec (Hulth, 2003) (`docid: 2040.abstr`, `https://doi.org/10.1023/A:1015266930361`).

pour améliorer les performances d'applications en aval comme le résumé automatique de texte (Zha, 2002; Wan et al., 2007; Litvak and Last, 2008; Qazvinian et al., 2010; Liu et al., 2021), la classification de documents (Hammouda et al., 2005; Hulth and Megyesi, 2006; Han et al., 2007), l'analyse de sentiments (Berend, 2011; Glaser and Schütze, 2012), la recommandation de documents (Ferrara et al., 2011; Collins and Beel, 2019) ou les systèmes de question-réponse (Subramanian et al., 2018; Yang et al., 2019a; Lee et al., 2021).

Le coût prohibitif [1] de l'annotation manuelle et son impractibilité à grande échelle font que seule une fraction des documents présents dans les bibliothèques numériques sont pourvus de mots-clés. De plus, il s'agit le plus souvent de mots-clés renseignés par les auteurs des documents eux-mêmes, sans qu'ils aient été formés pour cela. Il en résulte une annotation fragmentaire, subjective et irrégulière des documents (Strader, 2009). C'est pour lutter contre cet écueil que, dès les années 1990, de nombreux chercheurs se sont intéressés à automatiser la production des mots-clés (Steier and Belew, 1993; Krulwich and Burkey, 1996). Deux courants de recherche s'opposent alors, avec d'un coté l'assignation de mots-clés, i. e. la sélection comme mots-clés des entrées d'un vocabulaire contrôlé (Leung and Kan, 1997), et de l'autre l'extraction de mots-clés, i. e. la sélection des mots-clés directement dans le texte source (Witten et al., 1999). Par la suite, l'assignation de mots-clés sera progressivement délaissée au profit de l'extraction, plus simple à mettre en œuvre et surtout n'étant pas conditionnée par l'existence d'un vocabulaire contrôlé.

La carence de documents annotés en mots-clés a longtemps contraint les chercheurs a privilégier les approches par apprentissage non supervisé. Ainsi, les premiers travaux sur l'extraction automatique de mots-clés reposent essentiellement sur des critères statistiques pour apprécier l'importance de mots-clés candidats identifiés à l'aide de patrons grammaticaux (Hulth, 2003; Tomokiyo and Hurst, 2003; Liu et al., 2009, 2011). Parmi ces critères, certains se détachent par leur efficience comme la position de la première occurrence dans le document ou le poids TF×IDF, et font écho à ceux utilisés dans d'autres applications analogues comme le résumé automatique (Hovy and Lin, 1998) ou l'extraction terminologique (Kim et al., 2009). Une seconde série de travaux, détaillée en §1.3 et dans laquelle nous situons une partie de nos contributions, s'appuient sur les méthodes de graphes pour trier les mots-clés candidats par ordre d'importance (Zha, 2002; Mihalcea and Tarau, 2004; Wan et al., 2007; Wan and Xiao, 2008a).

En parallèle, d'autres travaux ont exploré la piste de l'apprentissage supervisé en formalisant l'extraction automatique de mots-clés comme une tâche de classification binaire (i. e. mot-clé ou non mot-clé) (Turney, 1999; Witten et al., 1999; Hulth, 2003; Nguyen and Kan, 2007; Lopez and Romary, 2010). Ces approches utilisent des algorithmes traditionnels de classification (e. g. classification naïve bayésienne) qui nécessitent peu de données d'entraînement mais qui supposent la définition

1. Rares sont les données chiffrées à ce sujet, mais le coût moyen de l'indexation manuelle d'un article dans PubMed serait estimé à $10, https://lhncbc.nlm.nih.gov/ii/information/about.html

« manuelle » et experte de traits caractéristiques (*features*). Leur niveau de précision reste malgré tout comparable à celui des meilleures approches non supervisées (Kim et al., 2010; Hasan and Ng, 2014; Gallina et al., 2020).

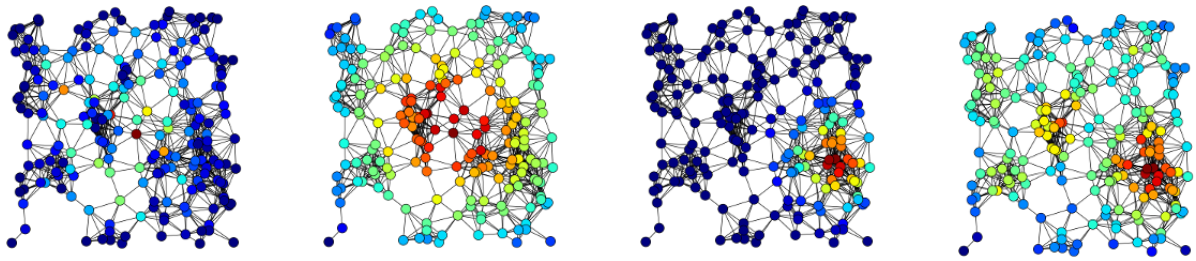
L'avènement des architectures neuronales profondes (Goodfellow et al., 2016), couplée à la récente mise à disposition de jeux de données de grande dimension (Meng et al., 2017; Gallina et al., 2019), a initié un nouveau courant d'approches construites autour des modèles séquence-à-séquence (Sutskever et al., 2014; Bahdanau et al., 2014). Ces approches neuronales ont permis de franchir une nouvelle étape dans la précision des mots-clés extraits (Meng et al., 2017; Chen et al., 2018; Ye and Wang, 2018), et de rendre leur application pertinente en pratique dans un contexte d'indexation documentaire (Boudin et al., 2020b). À noter que ces approches permettent aussi de « générer » des mots-clés, i. e. produire des mots-clés qui n'apparaissent pas dans le texte source, ce qui n'était jusque-là possible que dans le cadre de l'assignation de mots-clés (Medelyan and Witten, 2006, 2008). Ces derniers sont particulièrement utiles pour l'indexation documentaire puisqu'ils étendent le contenu des documents, mais ils sont aussi bien plus difficiles à produire car l'espace de recherche n'est plus borné aux sous-séquences de mots du document source.

1.3. Méthodes de graphes pour l'ordonnancement de texte

La théorie des graphes est une discipline des mathématiques discrètes dont l'objet d'étude est le *graphe*, structure mathématique utilisée pour modéliser des relations entre paires d'objets. Depuis sa création, qui remonte à la résolution du problème des sept ponts de Königsberg par Euler (1741), la théorie des graphes est régulièrement évoquée pour résoudre des problèmes dans diverses disciplines (e. g. transports, réseaux de télécommunications). C'est également le cas dans le TAL, où la structure en graphe permet de modéliser naturellement les relations entre les différentes unités de la langue, qu'il s'agisse de mots, de syntagmes ou de phrases entières. Ainsi, cette concordance conceptuelle entre graphe et langue naturelle se manifeste dans un grand nombre de travaux. Citons par exemple les travaux sur les ressources lexicales avec leurs réseaux sémantiques (Collins and Loftus, 1975; Miller, 1992; Baker et al., 1998), ceux sur l'analyse syntaxique avec leurs arbres de constituants et de dépendances (Klein and Manning, 2003; Chen and Manning, 2014), ou ceux sur la structure rhétorique du texte avec leurs graphes d'unités discursives élémentaires (Mann and Thompson, 1988).

Représenter le texte sous la forme d'un graphe permet d'accéder à un large éventail d'algorithmes et de modèles issus de la théorie des graphes. Ainsi, les algorithmes permettant de résoudre efficacement des problèmes de *plus court chemin* ou de *coloration de graphe* sont devenus la solution *de facto* pour plusieurs applications du TAL, comme pour l'analyse syntaxique (Maruyama, 1990; Huang and Chiang, 2005), l'extraction de relations (Bunescu and Mooney, 2005), la compression de phrases (Filippova, 2010) ou la traduction automatique (Ueffing et al., 2002). Dans nos travaux, nous nous sommes intéressé à une

autre famille d'algorithmes, celle des méthodes *d'ordonnancement de sommets*. Ces méthodes cherchent à estimer l'importance des sommets dans un graphe par l'intermédiaire d'indicateurs de centralité, qui assignent un rang à chaque sommet en fonction de leur position dans le graphe (Borgatti and Everett, 2006). Chaque indicateur donne une interprétation différente de la notion d'importance des sommets d'un graphe (voir Figure 1.3), et le choix du meilleur indicateur dépendra donc de l'application visée.



(a) **(b)** **(c)**

Figure 1.3. – Illustration des indicateurs de centralité intermédiaire (1.3a), de proximité (1.3b), de vecteur propre (1.3c) et de degré (1.3b) reprises de `https://commons.wikimedia.org/w/index.php?curid=39064835` (Tapiocozzo - CC BY-SA 4.0).

En RI, l'algorithme d'ordonnancement de sommets le plus connu est sans doute PageRank (Brin and Page, 1998), à l'origine utilisé par le moteur de recherche Google pour le classement des pages web. PageRank a ensuite été largement utilisé en TAL pour résoudre des problèmes d'ordonnancement de texte, notamment pour le résumé automatique (Mihalcea, 2004; Erkan and Radev, 2004; Ponza et al., 2018), l'extraction de mots-clés (Mihalcea and Tarau, 2004; Blanco and Lioma, 2007), et la désambiguïsation lexicale (Mihalcea et al., 2004; Agirre and Soroa, 2009; Scozzafava et al., 2020). Le principe utilisé dans les travaux susmentionnés est toujours le même : représenter le texte sous la forme d'un graphe où les sommets sont les unités à ordonner et les arêtes sont les relations qui les unissent. Cette représentation sera également reprise dans de nombreux travaux exploitant les réseaux de neurones sur graphes (Scarselli et al., 2009), et plus particulièrement pour la recommandation de documents (He et al., 2020; Wu et al., 2021b) et les systèmes de question-réponse (Huang and Yang, 2021; Fei et al., 2021).

1.4. Organisation du manuscrit

La suite de ce manuscrit est organisée en cinq chapitres. Le Chapitre 2 présente nos travaux sur les méthodes de graphe pour l'indexation par mots-clés. Le Chapitre 3 s'intéresse à l'évaluation automatique des mots-clés et décrit nos travaux sur l'évaluation indirecte au travers de tâches applicatives. Le Chapitre 4 revient sur plusieurs de nos travaux mêlant techniques du TAL et recherche d'information. Le Chapitre 5 dresse un panorama de nos travaux sur la construction de ressources langagières, le développement d'outils logiciels et leur valorisation dans la communauté scientifique. Dans le Cha-

pitre 6, nous partageons nos réflexions et proposons quelques pistes de recherche qui nous semblent prometteuses. La Figure 1.4 donne une vue d'ensemble de notre production scientifique et positionne chronologiquement les publications à partir desquelles ont été élaborés chaque chapitre.

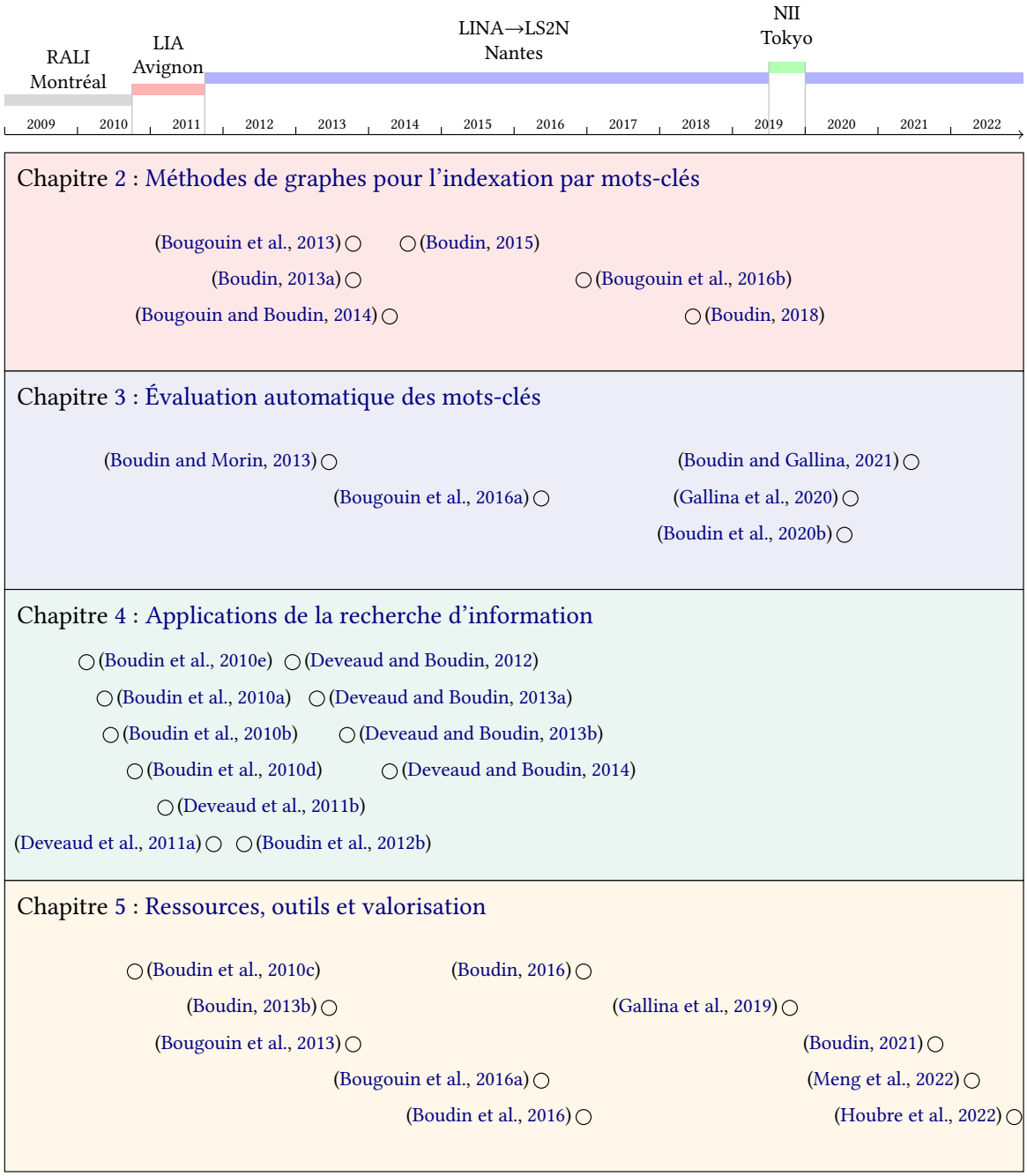


Figure 1.4. – Distribution chronologique de nos publications pour chaque chapitre du manuscrit.

I am wise enough to know that there are some perils from which a man must flee.

J.R.R. Tolkien, The Two Towers

2

Méthodes de graphes pour l'indexation par mots-clés

Sommaire

Ce chapitre aborde nos contributions sur les méthodes de graphes pour la production automatique de mots-clés. C'est TextRank, un modèle d'ordonnancement d'unités textuelles proposé par Mihalcea and Tarau (2004), qui sera le point de départ de nos réflexions. Dans ce modèle, un document est représenté sous la forme d'un graphe où les sommets sont les unités à ordonner et les arêtes sont les relations qui les unissent. L'ordonnancement est obtenu par l'exécution d'un algorithme dérivé de PageRank (Brin and Page, 1998), qui repose sur le principe de la marche aléatoire sur le graphe et attribut un score à chaque sommet selon l'hypothèse qu'un sommet est important si beaucoup d'autres sommets y sont reliés, ou si les sommets qui le relient sont importants.

De façon formelle, soit le graphe $G = (V, E)$ où V est l'ensemble des sommets, et $E \subseteq V^2$ est l'ensemble des arêtes. Soit $N(v_i) = \{v_j \in V : (v_i, v_j) \in E\}$ l'ensemble des sommets adjacents (voisinage) de v_i. Le score d'importance d'un sommet $S(v_i)$ est calculé par la méthode de la puissance itérée (Mises and Pollaczek-Geiringer, 1929) à partir de l'équation (2.1) donnée ci-dessous.

$$S(v_i) = (1-d) + d \cdot \sum_{v_j \in N(v_i)} \frac{S(v_j)}{|N(v_j)|} \tag{2.1}$$

où $d \in [0, 1]$ est un paramètre d'atténuation qui représente la probabilité de sauter aléatoirement d'un sommet à un autre du graphe. Partant de valeurs arbitraires attribuées à chaque sommet, le calcul des

scores est répété jusqu'à convergence, i. e. lorsque l'écart des scores entre deux itérations est inférieur à un seuil ϵ donné.

L'approche initiale de TextRank pour l'extraction de mots-clés s'appuie sur une représentation en graphe de cooccurrences de mots (voir Figure 2.1). Plus spécifiquement, les mots correspondant à des noms et des adjectifs sont identifiés dans le texte et représentent les sommets du graphe. Deux sommets sont reliés par une arête si les mots qui s'y rapportent cooccurrent dans une fenêtre de N mots. Une fois ordonnés selon l'équation (2.1), les mots les plus importants sont marqués dans le texte, et les séquences contiguës de ces mots sont sélectionnées pour produire les mots-clés.

Compatibility of systems of linear constraints over the set of natural numbers.

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.

mots-clés : linear constraints – linear diophantine equations – minimal generating sets – non-strict inequations – set of natural numbers – strict inequations – upper bounds

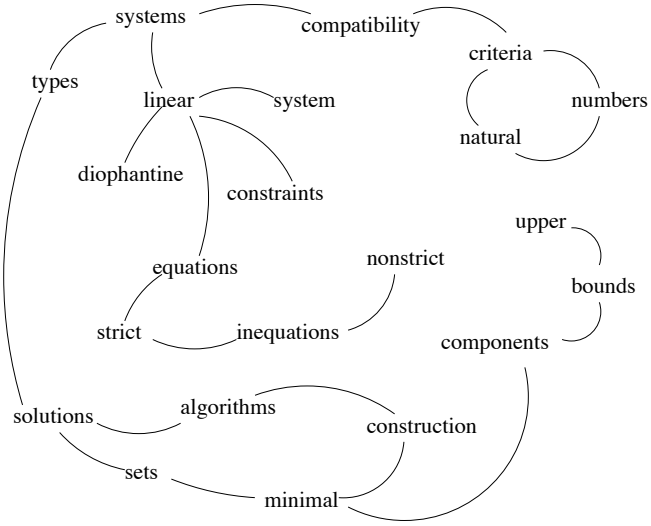


Figure 2.1. – Exemple de document de la collection de test Inspec (Hulth, 2003) (`docid: 1939.abstr`) et de sa représentation en graphe de cooccurrences de mots reprises de l'article de TextRank (Mihalcea and Tarau, 2004). Les mots-clés extraits par TextRank (i. e. séquences contiguës des mots les plus importants) sont marqués en rouge dans le texte.

Malgré son apparente simplicité, cette approche produit des résultats comparables à ceux de l'état de l'art de l'époque. TextRank présente néanmoins certaines limitations dont les plus prégnantes sont l'ordonnancement indirect des mots-clés (i. e. en fonction de l'importance des mots qu'ils contiennent) et

la sélection indépendante des mots-clés (i. e. sans tenir compte de la redondance ou de la représentativité de l'ensemble choisi). La suite de ce chapitre synthétise les travaux que nous avons menés pour lever ces limitations. Nous y décrivons trois modèles d'extraction de mots-clés (§2.1) et présentons une analyse empirique comparative de leur performance (§2.2).

2.1. Modèles proposés

Nous introduisons tout d'abord la notion de *sujet* qui est la clé de voûte des modèles d'extraction de mots-clés que nous avons proposé. Un sujet est définit comme l'ensemble des mots-clés candidats (i. e. les syntagmes nominaux) faisant référence à un même concept dans un document (voir Figure 2.2). Il s'agit d'une définition proche de celles des mentions coréférentielles en résolution de coréférence (Cardie and Wagstaff, 1999; Soon et al., 2001) mais simplifiée car n'incluant pas les mentions pronominales. L'hypothèse que nous formulons ici est la suivante : les mots-clés d'un document correspondent aux termes employés pour désigner les principaux sujets qui y sont abordés.

Inverse problems for a mathematical model of ion exchange in a compressible ion exchanger

A mathematical model of ion exchange is considered, allowing for ion exchanger compression in the process of ion exchange. Two inverse problems are investigated for this model, unique solvability is proved, and numerical solution methods are proposed. The efficiency of the proposed methods is demonstrated by a numerical experiment.

sujets : [ion exchange, compressible ion exchanger, ion exchanger compression] [mathematical model, model] [numerical solution methods, methods] [process] [inverse problems] [unique solvability] [efficiency] [numerical experiment]

Figure 2.2. – Exemple de document (présenté plus en détails dans la Figure 1.2) et d'un regroupement des mots-clés candidats en sujets par similarité lexicale.

2.1.1. TopicRank

Notre premier modèle, TopicRank (Bougouin et al., 2013), s'appuie sur une représentation en graphe pour classer les sujets d'un document par ordre d'importance, et sélectionne un ensemble de mots-clés représentatif parmi les sujets les plus importants. Plus précisément, un document est représenté sous la forme d'un graphe valué, dans lequel les sommets sont les sujets et les arêtes quantifient le lien sémantique qui les unit (voir Figure 2.3). Cette représentation « densifiée » capture plus finement la sémantique du document et intègre, de par le regroupement des mot-clés candidats, un moyen implicite de contrôler la redondance des mots-clés extraits. Pour rendre notre modèle généralisable (à différents

domaines, à d'autres langues), plusieurs choix méthodologiques ont été établis : 1) regrouper les mots-clés candidats en sujets par similarité lexicale, 2) délaisser le paramètre de fenêtre de cooccurrence au profit d'une pondération des arêtes selon la distance entre les mots-clés, et 3) utiliser la première position comme critère de sélection des mots-clés représentatifs de chaque sujet. Une étude de l'impact de ces différents choix sur les performances du modèle a été réalisée dans (Bougouin and Boudin, 2014).

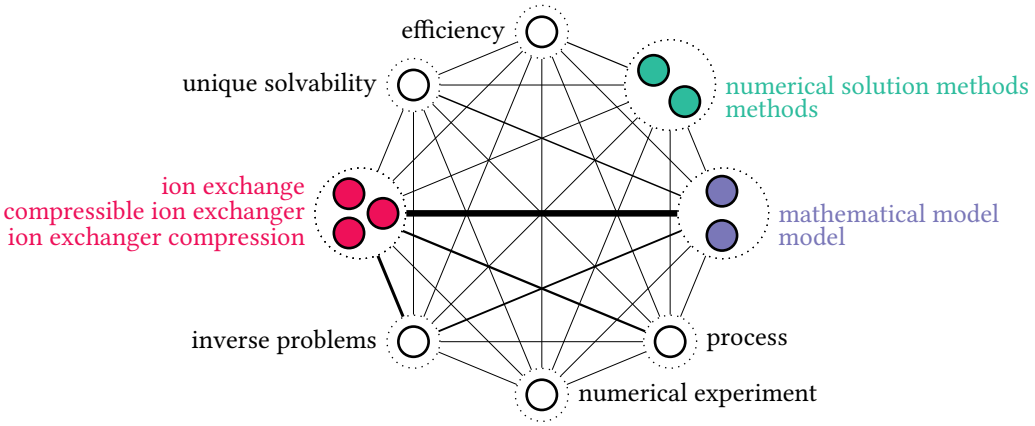


Figure 2.3. – Exemple de représentation en graphe de sujets du document présenté dans la Figure 2.2.

De façon formelle, soit w_{ij} le poids de l'arête entre les sommets v_i et v_j calculé à partir de l'équation (2.2) donnée ci-dessous.

$$w_{ij} = \sum_{c_i \in t_i} \sum_{c_j \in t_j} dist(c_i, c_j) \tag{2.2}$$

$$dist(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|} \tag{2.3}$$

où t_i est l'ensemble des mots-clés candidats du sujet représenté par le sommet v_i et $pos(c_i)$ correspond aux positions de décalage du candidat c_i dans le document. Le score d'importance d'un sommet $S(v_i)$ est obtenu selon l'équation (2.4) donnée ci-dessous, qui modifie l'équation (2.1) pour prendre en compte la pondération des arêtes.

$$S(v_i) = (1 - d) + d \cdot \sum_{v_j \in N(v_i)} \frac{w_{ij} \cdot S(v_j)}{\sum_{v_k \in N(v_j)} w_{jk}} \tag{2.4}$$

2.1.2. TopicCoRank

Notre second modèle, TopicCoRank (Bougouin et al., 2016b), est une adaptation faiblement supervisée de TopicRank pour la génération de mots-clés, i. e. la production de mots-clés n'apparaissant pas nécessairement dans le document source. Ce modèle s'appuie sur l'unification de deux représentations

en graphe (voir Figure 2.4), celle du document (graphe de sujets tel que présenté en Figure 2.3) et celle du domaine (graphe des mots-clés associés aux documents du même domaine). Il opère ensuite un ordonnancement conjoint pour y identifier les sujets et les mots-clés du domaine les plus importants vis-à-vis du document. Les relations (connexions) entre les sujets et les mots-clés du domaine établies lors de l'unification des deux graphes jouent ici le rôle de catalyseur, l'importance des uns entraînant celle des autres. La singularité de TopicCoRank réside dans l'acquisition automatique de connaissances du domaine à partir de documents annotés en mots-clés, qui se substituent aux thésaurus sur lesquels s'appuient les travaux précédents (Medelyan and Witten, 2008), et qui s'intègrent de manière transparente et contrôlable dans le modèle.

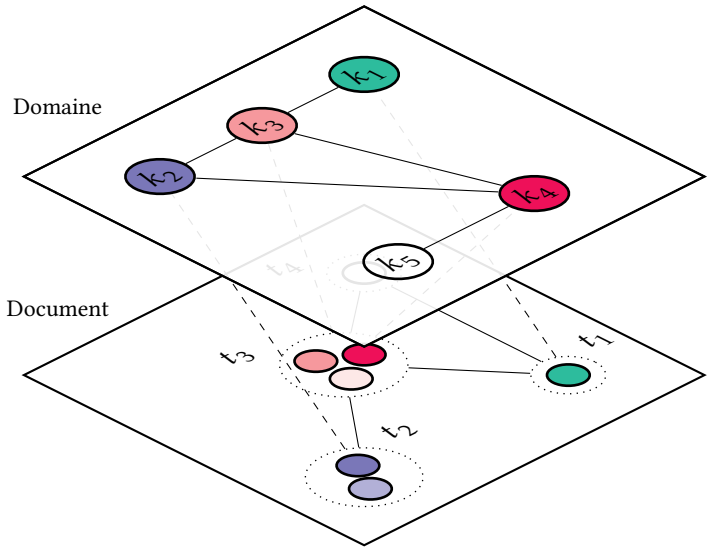


Figure 2.4. – Illustration du modèle TopicCoRank qui unifie les représentations en graphe du domaine (mots-clés de référence) et du document (sujets).

De façon formelle, soit le graphe unifié $G = (V, E)$ où $V = \{T \cup K\}$ est l'ensemble des sommets représentants les sujets $T = \{t_1, t_2, ..., t_n\}$ et les mots-clés du domaine $K = \{k_1, k_2, ..., k_m\}$, et $E = \{E_{in} \cup E_{ext}\}$ est l'ensemble des arêtes qui connectent deux sommets de même nature ($E_{in} \subseteq T^2 \cup K^2$) ou de nature différente ($E_{ext} \subseteq T \times K$). Deux sujets sont reliés si leurs candidats cooccurrent dans la même phrase (t_i —— t_j dans la Figure 2.4, $(t_i, t_j) \in E_{in}$), et deux mots-clés du domaine sont reliés si ils ont été associés au même document dans la collection d'entraînement (k_i —— k_j dans la Figure 2.4, $(k_i, k_j) \in E_{in}$). Par soucis d'homogénéité dans le graphe unifié, les arêtes sont pondérées selon la fréquence de cooccurrence des sujets ou d'association des mots-clés du domaine. Pour unifier les deux graphes, les mots-clés du domaine sont reliés aux sujets dans lesquels ils apparaissent (k_i - - - t_j dans la Figure 2.4, $(k_i, t_j) \in E_{ext}$). Le score d'importance d'un sommet $S(v_i)$ est obtenu selon l'équation (2.5) donnée ci-dessous, qui modifie l'équation (2.4) pour paramétriser la dynamique entre la recommandation interne R_{in} (i. e. dans chacun des deux graphes, équation (2.6)) et la recommandation

externe R_{ext} (i. e. entre les deux graphes, équation (2.7)) dans l'ordonnancement.

$$S(v_i) = (1 - \lambda) \cdot R_{ext}(v_i) + \lambda \cdot R_{in}(v_i) \tag{2.5}$$

$$R_{in}(v_i) = \sum_{v_j \in \overline{N}(v_i)} \frac{w_{ij} \cdot S(v_j)}{\sum_{v_k \in \overline{N}(v_j)} w_{jk}} \tag{2.6}$$

$$R_{ext}(v_i) = \sum_{v_j \in \ddot{N}(v_i)} \frac{S(v_j)}{|\ddot{N}(v_j)|} \tag{2.7}$$

où $\overline{N}(v_i) = \{v_j \in V : (v_i, v_j) \in E_{in}\}$ est l'ensemble des sommets connectés à v_i par une arête interne, et $\ddot{N}(v_i) = \{v_j \in V : (v_i, v_j) \in E_{ext}\}$ est l'ensemble des sommets connectés à v_i par une arête externe. Le paramètre $\lambda \in [0, 1]$ contrôle l'équilibre entre recommandation interne et externe.

2.1.3. MultipartiteRank

Notre troisième modèle, MultipartiteRank (Boudin, 2018), étend et améliore TopicRank en intervenant sur deux aspects interdépendants qui sont la représentation des sujets dans le graphe et la sélection des mots-clés représentatifs. Nous nous appuyons pour cela sur une structure en graphe multipartite (également appelé k-partite) dans laquelle les sommets, qui représentent les mots-clés candidats, sont partitionnés en k ensembles disjoints, chacun correspondant à un sujet différent (voir Figure 2.5). Cette structure particulière permet à l'algorithme d'ordonnancement d'agir directement sur mots-clés candidats tout en tirant profit du partitionnement pour distribuer l'importance entre les sujets. Autrement dit, l'importance accumulée par les sommets de chaque sujet est redirigée vers les sommets de sujets connexes, mettant en œuvre une relation de renforcement mutuel entre les sujets importants et les mots-clés représentatifs.

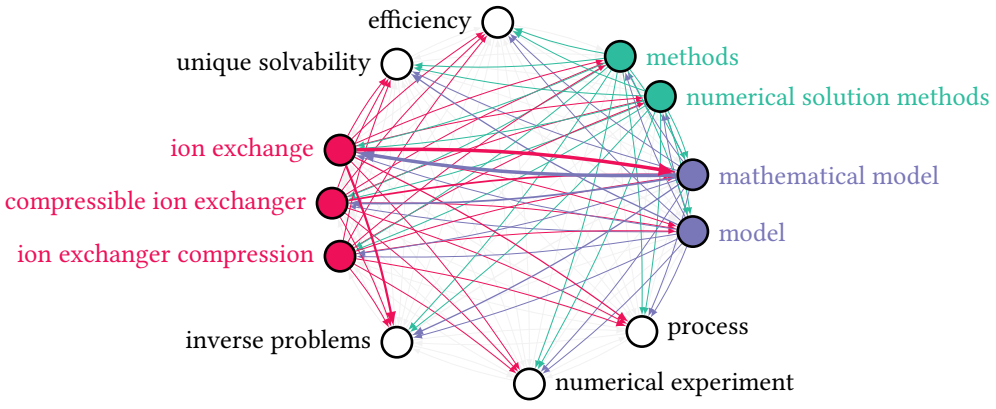


Figure 2.5. – Exemple de représentation en graphe multipartite du document présenté dans la Figure 2.2.

De façon formelle, MultipartiteRank s'appuie sur un graphe multipartite valué et orienté, dans lequel les sommets sont les mots-clés candidats et les arêtes, qui ne relient que les mots-clés de sujets différents, sont pondérées selon l'équation (2.3). Le score d'importance d'un sommet $S(v_i)$ est obtenu selon l'équation (2.8) donnée ci-dessous, qui modifie l'équation (2.4) pour tenir compte de l'orientation des arêtes.

$$S(v_i) = (1-d) + d \cdot \sum_{v_j \in pred(v_i)} \frac{w_{i,j} \cdot S(v_j)}{\sum_{v_k \in succ(v_j)} w_{jk}} \tag{2.8}$$

où $succ(v_i) = \{v_j \in V : (v_i, v_j) \in E\}$ et $pred(v_i) = \{v_j \in V : (v_j, v_i) \in E\}$ sont les ensembles respectifs des sommets successeurs et prédécesseurs de v_i.

Une autre originalité de MultipartiteRank est l'introduction d'un mécanisme de contrôle implicite de la sélection des mots-clés représentatifs des sujets. Nous agissons pour cela sur la pondération des arêtes des sommets que l'on veut promouvoir selon des critères supplémentaires. Plus précisément, le poids des arêtes entrantes du mot-clé candidat que l'on veut prioriser est augmenté en fonction du poids des arêtes sortantes des autres mots-clés candidats du sujet (voir Figure 2.6). À l'instar de TopicRank, notre modèle encourage la sélection des mots-clés candidats apparaissant au début du document en repondérant leurs arêtes entrantes selon l'équation (2.9) ci dessous.

$$w_{ij} = w_{ij} + \alpha \cdot e^{\left(\frac{1}{p_i}\right)} \cdot \sum_{v_k \in T(c_j) \setminus \{v_j\}} w_{ki} \tag{2.9}$$

où $T(c_j)$ est l'ensemble des sommets représentant des mots-clés candidats du même sujet que celui du mot-clé candidat c_j, et α est un paramètre qui détermine la force de l'ajustement des poids. À noter que ce mécanisme d'ajustement des poids peut être adapté à d'autres critères de sélection, comme la priorisation des mots-clés candidats correspondant aux entrées d'un vocabulaire contrôlé.

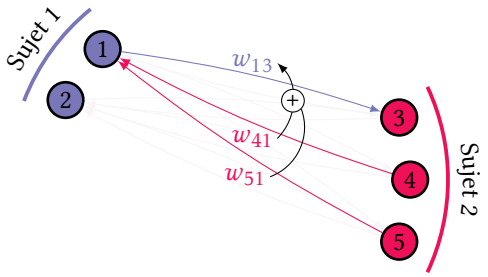


Figure 2.6. – Illustration du mécanisme d'ajustement des poids du graphe. Ici le sommet 3 est promu par l'augmentation du poids de son arête entrante en fonction des poids des arêtes sortantes des sommets 4 et 5.

2.2. Résultats

La Table 2.1 présente une partie des résultats obtenus par nos modèles en comparaison avec plusieurs *baselines* sur quatre ensembles de données aux propriétés différentes. Les mesures d'évaluation reportées sont la f-mesure au top-10 (F_1@10) et la moyenne des précisions moyennes (mAP), calculées par correspondance stricte entre les mots-clés produits automatiquement et les mots-clés de référence. Dans un souci de comparabilité des résultats, tous les modèles présentés reposent sur une chaîne de traitement unifiée.[1]

		Anglais				*Français*			
		Textes pleins		*Notices bibliographiques*				*Articles de presse*	
	Modèle	**SemEval**		**Inspec**		**TermITH**		**Wikinews**	
		F_1@10	mAP	F_1@10	mAP	F_1@10	mAP	F_1@10	mAP
①	TextRank	7.6	-	11.9	-	10.3	-	14.8	-
①	SingleRank	4.0	2.8	35.6	32.3	12.9	7.9	23.8	19.9
①	TopicRank	12.6	7.2	29.5	24.8	14.6	9.8	33.9	28.0
②	TopicCoRank	13.2	8.3	24.2	19.6	19.5*	17.2*	28.9*	24.2*
③	PositionRank	6.8	4.1	34.2	32.2	15.3	9.4	33.7	28.4
③	Topical PageRank	4.2	2.9	35.3	32.4	-	-	-	-
③	MultipartiteRank	14.3	10.6	30.5	29.0	16.0	10.9	36.3	31.7

Table 2.1. – Résultats en matière de F_1@10 et de mAP pour nos modèles et différentes *baselines*. * indique que les scores ont été calculés par validation croisée d'un contre tous (*leave-one-out cross-validation*). Les scores soulignés indiquent une amélioration significative (t.test < 0.05).

Dans une première série de résultats (voir ① dans la Table 2.1), nous comparons les scores de TopicRank avec ceux obtenus par TextRank et SingleRank (Wan and Xiao, 2008a), qui étends TextRank en pondérant les arêtes selon la fréquence de cooccurrence des mots et ordonne les mots-clés candidats en fonction de la somme des scores d'importance des mots qu'ils contiennent. TopicRank obtient les meilleurs résultats sur trois des quatre ensembles de données considérés, ce qui valide l'intérêt de l'ordonnancement des sujets du document pour l'extraction des mots-clés.

TopicCoRank, notre modèle faiblement supervisé de génération de mots-clés, donne des scores plus élevés sur deux des quatre ensembles de données (voir ② dans la Table 2.1). Ces résultats confirment l'apport positif des connaissances du domaine sur l'ordonnancement des mots-clés candidats, et

1. Nous utilisons l'outil Stanford CoreNLP (Manning et al., 2014) pour le découpage du texte en phrases, la *tokenisation* et l'étiquetage grammatical. Les mots-clés candidats sont identifiés à partir des patrons grammaticaux `/Adj*Noun+/` et `/Noun+Adj*/` pour l'anglais et le français respectivement. Nous utilisons les modèles implémentés dans l'outil `pke` (Boudin, 2016) avec les paramètres par défaut.

représentent un premier pas réussi dans la transition des modèles extractifs vers des modèles génératifs. Il est important de noter que les scores de TopicCoRank surpassent significativement ceux de tous les autres modèles sur l'ensemble de données TermITH, seul ensemble où les mots-clés de référence ont été annotés par des indexeurs professionnels. Ce résultat met en exergue le rôle prépondérant que jouent les annotations expertes dans l'apprentissage et l'évaluation des modèles.

La dernière série de résultats (voir (3) dans la Table 2.1) met en comparaison MultipartiteRank avec deux modèles de l'état de l'art construits autour de l'algorithme de PageRank biaisé (Haveliwala, 2002). Le premier modèle, PositionRank (Florescu and Caragea, 2017), combine position et fréquence des mots pour infléchir l'ordonnancement des mots-clés, tandis que le second modèle, Topical PageRank (Liu et al., 2010) ici dans sa version modifiée par (Sterckx et al., 2015), utilise les *concepts* calculés via Latent Dirichlet Allocation (LDA) (Blei et al., 2003). MultipartiteRank produit les meilleurs résultats sur trois des quatre ensembles de données considérés, et fonctionne particulièrement bien sur les documents longs (SemEval), pourtant reconnus pour être difficile à traiter (Hasan and Ng, 2014).

2.3. Discussion

Dans ce chapitre, nous avons présenté en détails trois méthodes de graphes pour l'indexation par mots-clés bâties autour de la notion de regroupement des mots-clés candidats en *sujets*. La simplicité de mise en œuvre des méthodes proposées, associée aux bons résultats constatés sur de nombreux ensembles de données font qu'elles sont régulièrement utilisées en qualité de *baseline* par la communauté, comme par exemple dans (Florescu and Caragea, 2017; Bennani-Smires et al., 2018; Çano and Bojar, 2019; Santosh et al., 2020; Patel and Caragea, 2021). Notre méthode TopicCoRank annonce quant à elle la transition vers les modèles génératifs, qui sera ensuite largement accélérée par l'avènement des méthodes par apprentissage profond (Meng et al., 2017).

Parmi les différentes pistes d'amélioration possibles pour nos méthodes, une apparaît particulièrement évidente : remplacer le regroupement par similarité lexicale des mots-clés candidats en sujets par une technique moins naïve et plus robuste. Pourtant, les différentes tentatives que nous avons entreprises dans cette direction, comme le recours à des mesures de similarité sémantique ou à des algorithmes de regroupement supervisés, se sont avérées infructueuses. Parmi les éléments bloquants que nous avons identifié, deux ressortent : la sélection figée des mots-clés candidats (i. e. par patron grammatical) qui omet une quantité importante de candidats potentiels, et sa décorrélation d'avec l'algorithme de regroupement qui ne peut exploiter au mieux le contexte dans lesquels les candidats sont ancrés. Une piste intéressante pour lever ces obstacles pourrait être l'adaptation de méthodes d'extraction d'entités-relations, et en particulier celles non supervisées qui, à la fois, offrent un niveau de précision satisfaisant et sont suffisamment flexibles pour répondre à nos besoins (Tran et al., 2020; Brody et al., 2021).

Aussi, certains aspects des méthodes de graphes pour l'indexation par mots-clés restent pratiquement inexplorés. En particulier, rares sont les travaux qui utilisent d'autres algorithmes que PageRank pour ordonner les sommets. Nous avons pourtant montré que des indicateurs de centralité, comme la centralité de degré ou de proximité (*closeness centrality*), pouvaient donner des résultats aussi bons, voire meilleurs (Boudin, 2013a). D'autres travaux vont dans ce sens (Schluter, 2014; Tixier et al., 2016) et, dans le cadre supervisé, les réseaux convolutifs sur graphes (*graph convolutional networks*) affichent des résultats probants (Prasad and Kan, 2019; Sun et al., 2019).

Toujours dans l'optique d'améliorer l'ordonnancement des sommets, plusieurs pistes de recherche semblent sous-explorées, comme par exemple l'utilisation d'algorithmes semi-supervisés (Alexandrescu and Kirchhoff, 2007; Rao and Yarowsky, 2009) ou la prise en considération de connaissances externes (Gao et al., 2011; Zhang et al., 2019). Concernant ce dernier point, il serait intéressant d'explorer d'avantage ce que peut apporter les approches automatisées de construction de connaissances, et notamment les approches neuronales non supervisées dont le niveau de performance ne cesse d'augmenter (Tran et al., 2020; Yuan and Eldardiry, 2021).

De manière transverse, l'efficacité des algorithmes d'ordonnancement en termes de coût calculatoire est un aspect rarement évoqué dans les travaux sur l'indexation par mots-clés. Pourtant, et même si ces algorithmes sont exécutés de façon hors-ligne, cet aspect mériterait d'être pris en compte et étudié tant les volumes de documents à traiter dans le cadre de l'indexation documentaire sont conséquents. Des initiatives existent en ce sens et constitueraient d'intéressantes pistes à explorer, comme l'utilisation de méthodes de calcul plus efficaces (Bahmani et al., 2010; Zhu et al., 2013; Mitliagkas et al., 2015) ou l'accélération des calculs sur processeur graphique (GPU) (Shi et al., 2019; Guo et al., 2017).

De même, très peu d'études se sont penchées sur la sélection de l'ensemble optimal de mots-clés, i. e. la combinaison minimale de mots-clés qui maximise la couverture informationnelle. Cette problématique n'est d'ailleurs pas limitée aux seules méthodes de graphes et va au delà de la question de la prédiction du nombre de mots-clés à inférer (Yuan et al., 2020). Nos modèles, de part un ordonnancement interdépendant des sommets du graphe, apportent un début de solution au problème de couverture informationnelle des mots-clés mais cela n'est évidemment pas suffisant. Des techniques d'optimisation combinatoire post-ordonnancement ont également été proposées pour lutter plus efficacement contre ce problème mais elles nécessitent toujours de fixer a priori la taille de l'ensemble de mots-clés (Ding et al., 2011; Boudin, 2015).

It's the job that's never started as takes longest to finish.

J.R.R. Tolkien, The Fellowship of the Ring

3

Évaluation automatique des mots-clés

Sommaire

Dans la littérature, le principal paradigme d'évaluation pour mesurer la qualité des mots-clés produits consiste à les comparer à une vérité terrain, i. e. à des mots clés de référence saisis manuellement, préférablement par des indexeurs professionnels. Les scores obtenus selon cette évaluation intrinsèque, qui s'appuie sur une seule et unique réponse correcte, ne reflètent pas le niveau réel de qualité des mots-clés produits (Medelyan and Witten, 2006), et surtout ne permettent pas de quantifier avec exactitude leur contribution aux applications en aval. Il faut ajouter à cela que les mots-clés de référence, qui pour des raisons de coût et de disponibilité sont presque toujours ceux renseignés par les auteurs des documents eux-mêmes, ne sont pas exempts d'incohérences et de lacunes. L'absence de consensus sur les paramètres expérimentaux entre les différentes études contribue à noircir davantage le tableau de l'évaluation intrinsèque, amoindrissant ainsi la portée des conclusions qui y sont avancées.

Ce chapitre présente les travaux que nous avons menés pour résoudre ces différents problèmes, en commençant par une analyse contrastive des évaluations manuelles et automatiques des mots-clés (§3.1). Nous décrivons nos expériences sur l'évaluation extrinsèque des mots-clés au travers de plusieurs cadres applicatifs (§3.2), et interprétons les résultats d'une évaluation intrinsèque comparative à grande échelle des modèles de l'état de l'art (§3.3).

3.1. Évaluation manuelle contre évaluation automatique

L'évaluation manuelle des méthodes d'indexation par mots-clés consiste à faire valider les ensembles de mots-clés proposés par un ou plusieurs indexeurs professionnels (Jones and Paynter, 2003). Parce qu'elle est coûteuse et peu reproductible, cette forme d'évaluation est systématiquement remplacée par une évaluation automatique. Évidemment, l'évaluation automatique des mots-clés est moins fiable, mais peu de travaux ont été entrepris pour mesurer l'écart qualitatif qui la sépare de l'évaluation manuelle. Pour répondre à cette question, nous avons élaboré un protocole d'évaluation manuelle en deux critères et comparé les résultats obtenus à ceux d'une évaluation automatique (Bougouin et al., 2016a). Les deux critères retenus, chacun noté sur une échelle de 0 à 2, mesurent respectivement la pertinence des mots-clés et la quantité d'information perdue par rapport aux mots-clés de référence. Ils sont décrits de façon formelle ci-dessous :

Pertinence : ce critère distingue les mot-clés pertinents (score de 2), des formes variantes (score de 1) et de ceux non pertinents (score de 0). À noter qu'une forme variante peut correspondre à un autre mot-clé, et sera marquée comme redondante.

Silence : ce critère quantifie la perte d'information par rapport aux mots-clés de référence. Il distingue si l'information perdue est capitale (score 2), secondaire (score 1) ou si il n'y a pas de perte d'information (score 0).

La Table 3.1 présente une partie des résultats des évaluations manuelles réalisées par des indexeurs professionnels sur l'ensemble de données TermITH.[1] Les sorties (top-10) de deux méthodes d'extraction automatique de mots-clés sont comparées : TF-IDF et TopicRank, notre méthode de graphes décrite en §2.1.1. Pour l'évaluation de la Pertinence, nous distinguons les formes variantes (score de 1) redondantes de celles qui ne le sont pas. Nous reportons également les scores de f-mesure (F_1@10) calculés automatiquement et manuellement [2] par rapport aux mots-clés de référence.

On note une contradiction manifeste entre les scores d'évaluation automatique (voir la colonne **Auto.** dans la Table 3.1) qui placent TF-IDF en premier, et ceux obtenus selon le protocole d'évaluation manuelle qui donnent un avantage systématique à TopicRank. De plus, l'évaluation manuelle des mots-clés par rapport aux mots-clés de référence (voir la colonne **Man.** dans la Table 3.1) atteste, avec un gain d'environ 20 points, du pessimisme de l'évaluation automatique. L'évaluation de la Pertinence des formes variantes montre que TopicRank produit très peu de mots-clés redondants (voir la colonne red. dans la Table 3.1), ce qui témoigne sous un angle différent de l'utilité du regroupement des mots-clés candidats en sujets pour réduire la redondance dans les sorties (§2.1).

1. Pour un ensemble plus large de méthodes d'indexation par mots-clés et une analyse approfondie des résultats, le lecteur est invité à consulter (Bougouin, 2015).

2. Les mots-clés avec un score de Pertinence de 2 sont considérés corrects, de même que ceux avec un score de 1 non redondants.

Modèle	Pertinence				Silence			Man.	Auto.
	0	1		2	0	1	2	F_1@10	F_1@10
		red.	¬red.						
TF-IDF	53.8	6.8	4.2	35.3	31.4	48.5	20.1	33.5	**13.9**
TopicRank	**56.3**	**0.9**	**5.7**	**37.1**	**35.0**	**48.3**	**16.8**	**36.3**	11.9

Table 3.1. – Taux de mots-clés par score $\in [0, 1, 2]$ pour les critères de Pertinence et de Silence dans les sorties (top-10) de chaque méthode. Les résultats en matière de F_1@10 calculés automatiquement (Auto.) et manuellement (Man.) par correspondance entre les mots-clés produits et les mots-clés de référence sont également reportés.

Ces résultats soulignent la fragilité potentielle de l'évaluation automatique et commandent d'utiliser d'autres stratégies pour mesurer l'efficacité des méthodes d'indexation par mots-clés. L'évaluation extrinsèque est une de ces stratégies, en particulier au moyen d'applications pour lesquelles l'adjonction de mots-clés a impact direct et mesurable sur les performances.

3.2. Évaluation extrinsèque

Cette section présente nos travaux sur l'évaluation extrinsèque des mots-clés au travers de deux cadres applicatifs : la recherche documentaire et le résumé automatique.

3.2.1. Évaluation dans un cadre de recherche documentaire

La recherche documentaire est l'application qui vient en premier à l'esprit lorsque l'on cherche à évaluer les mots-clés de manière extrinsèque. Pourtant, depuis la fin des années 90 et les travaux sur les moteurs de recherche interactifs (Gutwin et al., 1999; Jones and Staveley, 1999), aucune étude n'a, à notre connaissance, exploité la recherche de documents pour évaluer indirectement la qualité des mots-clés produits automatiquement.[3] Nous attribuons cette singularité au quasi-monopole des méthodes extractives qui, du point de vue de la recherche d'information, n'ont en théorie que peu d'utilité puisqu'elles ne font que modifier la fréquence des termes dans les documents.

Le tournant récemment opéré vers les méthodes génératives change fondamentalement la donne puisqu'elles sont capables de produire des mots-clés qui n'apparaissent pas dans le texte source. Les méthodes de génération de mots-clés s'apparentent donc d'avantage à des techniques d'expansion de documents (Tao et al., 2006; Efron et al., 2012) qui visent à tempérer la discordance de vocabulaire (*vocabulary mismatch*) (Furnas et al., 1987) entre la requête et les documents par l'ajout de mots en

3. Des travaux existent néanmoins dans le domaine biomédical qui évaluent l'impact des mots-clés assignés de manière automatique (i. e. issus de vocabulaires contrôlés comme le MeSH) sur la recherche documentaire (Névéol et al., 2006; Dinh et al., 2013).

relation sémantique avec le contenu du document (e. g. synonymes, hyperonymes). Pour mesurer et valider l'utilité de méthodes génératives pour la recherche documentaire, nous avons mené des expériences visant à comparer l'efficacité de différents systèmes de recherche d'information sur une collection de documents indexée avec et sans mots-clés (Boudin et al., 2020b).

Évaluation des méthodes génératives

La Table 3.2 présente les résultats obtenus par les systèmes de recherche d'information Okapi-BM25 (Robertson et al., 1999) et Query Likelihood (Ponte and Croft, 1998) sur la collection de test NTCIR-2 (Kando, 2001) à laquelle deux méthodes neuronales de génération de mots-clés de l'état de l'art ont été appliquées.[4] La première méthode, CopyRNN (Meng et al., 2017), s'appuie sur un modèle de séquence-à-séquence (Cho et al., 2014; Sutskever et al., 2014) avec mécanismes d'attention (Bahdanau et al., 2014) et de copie (Gu et al., 2016). La seconde méthode, CorrRNN (Chen et al., 2018), étend la première avec une mécanisme de couverture (Tu et al., 2016) qui augmente la diversité des mots-clés générés. À des fins de comparaisons, nous rapportons les résultats obtenus avec MultipartiteRank, notre méthode de graphes pour l'extraction de mots-clés décrite en §2.1.3. Nous rapportons également les résultats obtenus par l'adjonction d'une technique d'expansion de requêtes, ici RM3 (Abdul-Jaleel et al., 2004), pour des performances compétitives avec l'état de l'art (Yang et al., 2019b). La mesure d'évaluation reportée est la mAP calculée au 1000 premiers documents retournés.

	Index	**BM25**	**+RM3**	**QL**	**+RM3**	F_1@5
①	titre, résumé	29.2	31.9	29.0	31.5	-
	+ mots-clés (CopyRNN)	30.5	**34.3**	30.6	33.3	23.9
	+ mots-clés (CorrRNN)	30.3	33.2	29.8	31.4	22.3
	+ mots-clés (MultipartiteRank)	29.2	32.3	29.6	32.3	18.1
②	titre, résumé, mots-clés (auteurs)	31.4	35.2	30.6	33.0	-
	+ mots-clés (CopyRNN)	31.6	**36.5**	31.7	35.2	-
	+ mots-clés (CorrRNN)	31.4	35.8	31.1	33.7	-
	+ mots-clés (MultipartiteRank)	31.4	35.2	31.2	33.5	-

Table 3.2. – Résultats en matière de mAP des différents systèmes de recherche d'information en fonction de la configuration d'indexation utilisée. Les scores soulignés indiquent une amélioration significative (t.test < 0.05).

4. Nous utilisons l'implémentation de Okapi-BM25 de l'outil `anserini` (Yang et al., 2017) avec les paramètres par défaut (i. e. $k1 = 0.9$ et $b = 0.4$). Les deux méthodes génératives ont été implémentées en PyTorch (Paszke et al., 2017) dans l'outil AllenNLP (Gardner et al., 2018), et entraînées sur l'ensemble de données KP20k (Meng et al., 2017) avec les paramètres retenus par les auteurs. Le top-5 des mots-clés générés est utilisé pour étendre les documents, en écho au nombre moyen de mots-clés auteurs (4.8).

La première série de résultats (voir (1) dans la Table 3.2) atteste de l'intérêt des méthodes de génération de mots-clés pour la recherche documentaire puisqu'une amélioration significative des performances de tous les systèmes de recherche d'information est constatée. On note cependant que seul CopyRNN, qui obtient les meilleurs scores selon l'évaluation intrinsèque (voir la colonne F_1@5 dans la Table 3.2), permet une amélioration systématique des performances de recherche d'information. Ces résultats montrent que les méthodes neuronales de génération de mots-clés ont franchi le seuil de performance suffisant pour être exploitées en pratique dans un contexte d'indexation documentaire.

Plus intéressant encore, la deuxième série de résultats (voir (2) dans la Table 3.2) montre une amélioration, parfois significative, des performances des systèmes de recherche d'information lorsque les mots-clés produits par les méthodes automatiques sont associés aux mots-clés auteurs, ce qui traduit une complémentarité entre ces derniers. Il est également intéressant de voir que les gains de performance apportés par l'expansion des requêtes (voir les colonnes +RM3 dans la Table 3.2) et l'ajout de mots-clés sont additifs, suggérant qu'ils fournissent des signaux de pertinence différents mais complémentaires.

La supériorité supposée des méthodes génératives par rapport aux méthodes extractives n'est en revanche pas confirmée dans nos résultats puisque l'analyse des sorties de CopyRNN montre que 97% des mots-clés générés apparaissent dans le texte source. Le gain de performance observé sur les méthodes génératives ne vient donc pas de leur capacité à produire des mots-clés absents, mais plutôt d'une meilleure précision dans l'extraction des mots-clés. Cette constatation interroge sur un possible retour en grâce des méthodes extractives dès lors que leur précision est suffisante, mais surtout sur l'impact potentiel des mots-clés absents sur la recherche documentaire qui reste indéterminé.

Redéfinition des mots-clés absents

Nous avons réalisé des expériences pour mesurer précisément l'impact des mots-clés absents sur les performances des systèmes de recherche d'information (Boudin and Gallina, 2021). Ce travail se caractérise par une redéfinition de la notion de mot-clé absent au travers du prisme de l'indexation. La plupart des travaux en génération de mots-clés adoptent la définition donnée par (Meng et al., 2017), dans laquelle les mots-clés qui ne correspondent à aucune séquence contiguë de mots du texte source sont considérés comme absents. Du point de vue de la recherche d'information, où les mots pleins racinisés sont utilisés pour indexer les documents, cette définition est clairement inadaptée tel que le montre l'exemple de la Figure 3.1.

On constate que pour certains mots-clés absents, tous les mots qu'ils contiennent apparaissent dans le texte source, ce qui brouille la frontière entre mots-clés absents et présents. Nous distinguons donc trois sous-catégories de mots-clés absents selon si tout, une partie ou aucun des mots qui les composent apparaissent dans le texte source. Aussi, nous proposons d'adopter un schéma de catégorisation plus

Study on the Structure of Index Data for Metasearch System

This paper proposes a new technique for Metasearch system, which is based on the grouping of both keywords and URLs. This technique enables metasearch systems to share information and to reflect the estimation of users' preference. With this system, users can search not only by their own keywords but by similarity of HTML documents. In this paper, we describe the principle of the grouping technique as well as the summary of the existing search systems.

mots-clés :

(présents) metasearch – search system

(absents) information sharing – information retrieval – user's behavior – retrieval support

Figure 3.1. – Exemple de document (notice bibliographique) de la collection de test NTCIR-2 (Kando, 2001) (`docid: gakkai-e-0001384947`).

fin, nommé PRMU et illustré ci-dessous avec des exemples pris dans la Figure 3.1. Contrairement à la classification binaire précédemment utilisée (i. e. présent ou absent), ce schéma de catégorisation établit une distinction effective entre les mots-clés qui étendent le document (i. e. *mixed* et *unseen*) et ceux qui surlignent les mots importants (i. e. *present* et *reordered*).

Present (*présent*) : mots-clés qui correspondent à des séquences de mots contiguës du texte source (e. g. *search system*).

Reordered (*réordonné*) : mots-clés dont tous les mots constitutifs apparaissent dans le texte source mais pas sous forme de séquences contiguës (e. g. *information sharing*).

Mixed (*mixte*) : mots-clés dont une partie seulement des mots constitutifs apparaît dans le texte source (e. g. *information retrieval*).

Unseen (*inexistant*) : mots-clés dont aucun des mots constitutifs n'apparaît dans le texte source (e. g. *retrieval support*).

La Table 3.3 présente les résultats obtenus par le système Okapi-BM25 sur la collection de test NTCIR-2 lorsque les mots-clés (auteurs) correspondant aux différentes catégories PRMU sont indexées. Individuellement, on note que les meilleurs scores sont atteints avec les mots-clés des catégories Mixed et Unseen, qui pourtant ne représentent qu'une fraction du nombre de mots-clés présents (voir (1) dans la Table 3.3). On retrouve le même comportement lorsque les mots-clés de ces deux catégories sont combinés, avec des scores systématiquement plus élevés que ceux obtenus avec les mots-clés des catégories Present et Reordered (voir (2) dans la Table 3.3). Les mots-clés présents restent néanmoins utiles puisque les meilleurs scores sont au final obtenus avec l'ensemble des catégories (voir (3) dans la Table 3.3). Ces résultats montrent que les mots-clés absents qui étendent le document ont un rôle majeur dans l'amélioration des performances des systèmes de recherche d'information. Ce constat, mis au regard de la faible proportion de nouveaux mots introduits par l'indexation des mots-clés (i. e. en

moyenne 20-25% des mots uniques des mots-clés n'apparaissent pas dans le texte source), met en lumière le potentiel encore non exploité des méthodes génératives dans la recherche documentaire. A noter que des expériences sur une seconde collection de test, omises ici pour des raisons de lisibilité mais qui figurent dans (Boudin and Gallina, 2021), consolident ces observations.

	index	BM25	+RM3	#nb (%)
	titre, résumé	29.6	32.8	-
1	+ mots-clés (Present)	30.7	33.5	2.9 (61.9%)
1	+ mots-clés (Reordered)	29.8	33.5	0.4 (8.1%)
1	+ mots-clés (Mixed)	30.8	**33.9**	0.8 (16.5%)
1	+ mots-clés (Unseen)	29.7	**33.9**	0.7 (13.5%)
2	+ mots-clés (P+R)	30.6	33.8	3.3 (70%)
2	+ mots-clés (M+U)	30.8	**34.3**	1.5 (30%)
3	+ mots-clés (R+M+U)	30.8	34.9	1.9 (38.1%)
3	+ mots-clés (P+R+M+U)	31.9	**35.5**	4.8 (100%)

Table 3.3. – Résultats en matière de mAP du système Okapi-BM25 en fonction de la configuration d'indexation utilisée. Le nombre moyen de mots-clés indexés (#nb) et sa proportion par rapport au nombre total de mots-clés sont reportés. Les scores soulignés indiquent une amélioration significative (t.test < 0.05).

Le schéma de catégorisation PRMU donne un nouvel angle d'évaluation pour mesurer la capacité des méthodes génératives à produire des mots-clés absents. La Figure 3.2 compare les distributions par catégorie des mots-clés générés avec celle des mots-clés auteurs sur la collection de test NTCIR-2. Il en ressort que les méthodes génératives considérées ne produisent quasiment pas de mots-clés absents qui étendent le document. Ce constat pointe du doigt la principale faiblesse de ces méthodes, qui en théorie sont capables de produire des mots-clés n'apparaissant pas dans le texte source mais qui dans la pratique ne font pas. Les résultats mettent en évidence la complexité de la génération par rapport

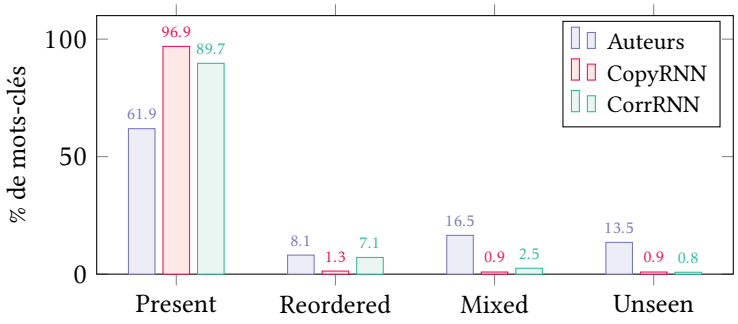


Figure 3.2. – Distributions par catégorie PRMU des mots-clés.

à l'extraction, et donnent une indication sur la direction à emprunter pour améliorer les méthodes neuronales actuelles.

3.2.2. Évaluation dans un cadre de résumé automatique

Les mots-clés donnent un aperçu synthétique du contenu d'un document, et constituent un résumé (très) condensé et dépourvu de syntaxe. Aussi, les méthodes de production automatique de mots-clés et celles de résumé automatique se rejoignent sur de nombreux aspects, notamment au niveau des modèles utilisés (Mihalcea and Tarau, 2004; Çano and Bojar, 2019). C'est pourquoi de nombreux travaux se sont intéressés à conjuguer ces deux problématiques, en particulier pour déterminer l'importance des unités textuelles qui composent les résumés (e. g. phrases, paragraphes) en fonction des mots-clés qu'elles contiennent (Zha, 2002; Wan et al., 2007; Qazvinian et al., 2010; Hong and Nenkova, 2014; Jadhav and Rajan, 2018). Dans cette même veine, nous nous sommes intéressé à la compression multi-phrases, tâche qui consiste à produire un résumé court (une phrase) à partir d'un ensemble de phrases connexes (voir la Figure 3.3), et nous avons étudié la possibilité de tirer parti des mot-clés pour produire des résumés plus informatifs.

Ensemble de phrases sources :
Last week the Secretary of State Ms. Clinton visited Chinese officials.
Hillary Clinton paid a visit to the People Republic of China on Monday.
The wife of a former U.S. president Bill Clinton Hillary Clinton visited China last Monday.
Hillary Clinton wanted to visit China last month but postponed her plans till Monday last week.

résumé : Hillary Clinton visited China on Monday.

Figure 3.3. – Exemple de résumé par compression multi-phrases repris de (Filippova, 2010).

Nous utilisons, pour point de départ, la méthode de compression multi-phrases proposée par (Filippova, 2010), qui s'appuie sur une représentation en graphe de cooccurrences de mots dans laquelle les plus courts chemins correspondent aux résumés. Pour améliorer l'informativité des résumés générés, i. e. la quantité d'information des phrases sources conservée dans le résumé, nous avons proposé un modèle de ré-ordonnancement qui cherche à maximiser l'importance des mots-clés qu'ils contiennent (Boudin and Morin, 2013). De façon formelle, soit C l'ensemble des résumés candidats générés par une méthode de compression multi-phrases $\mathcal{M}$, $\bar{S}(c)$ le score attribué au résumé c par $\mathcal{M}$, et $S(k)$ le score d'importance du mot-clé k. Le *meilleur* résumé c^* parmi l'ensemble C est sélectionné selon l'équation (3.1) donnée ci-dessous.

$$c^* = \arg\max_{c \in C} \frac{\bar{S}(c)}{l(c) \cdot \sum_{k \in c} S(k)} \tag{3.1}$$

où $l(c)$ est la longueur (en nombre de mots) du résumé c, et le score d'importance d'un mot-clé $S(k)$ est calculé par le modèle TopicRank (voir §2.1.1).

La Table 3.4 présente une partie des résultats obtenus par notre modèle de ré-ordonnancement par mots-clés sur la collection de test `LINA-msc` composée de 40 ensembles de phrases en français, chacun accompagné de trois résumés de référence. L'évaluation manuelle des résumés générés porte sur les critères de grammaticalité et d'informativité, chacun noté sur une échelle de 0 à 2 par trois annotateurs différents. Les métriques d'évaluation automatique standard ROUGE-1 (Lin, 2004) et BLEU (Papineni et al., 2002), qui comparent les résumés générés à ceux de référence, sont également reportées. On observe une amélioration significative du score d'informativité des résumés, qui est également visible sur les scores de l'évaluation automatique. Comme attendu, l'intérêt du ré-ordonnancement par mots-clés apparaît clairement avec un gain constant de 7 à 20% sans perte significative au niveau de la grammaticalité. Les mots-clés indiquent et surtout ordonnent les informations importantes de l'ensemble de phrases sources, et permettent d'identifier le résumé candidat qui offre le meilleur compromis entre longueur et informations conservées. Des résultats supplémentaires montrent que la métrique ROUGE-1 est corrélée au critère d'informativité (coefficient de Pearson $\rho = 0.59$), ce qui ouvre la possibilité d'évaluer l'impact de différents modèles de production de mots-clés sur la compression multi-phrases de manière automatique.

	évaluation manuelle		évaluation automatique		
Modèle	**Gram.**	**Info.**	ROUGE-1	BLEU	**TC**
(Filippova, 2010)	1.63	1.33	57.4	61.6	0.50
+ ré-ordonnancement	1.53 (−6%)	<u>1.60</u> (+20%)	<u>65.7</u> (+14%)	<u>65.8</u> (+7%)	0.58

TABLE 3.4. – Résultats du modèle proposée par (Filippova, 2010) avec et sans ré-ordonnancement par mots-clés. Les scores d'évaluation manuelle correspondent à la moyenne des scores des annotateurs, et TC indique le taux de compression des résumés. Les scores soulignés indiquent une amélioration significative (t.test < 0.01).

3.3. Évaluation intrinsèque à grande échelle

Il existe une grande disparité dans les paramètres expérimentaux utilisés pour évaluer les performances des modèles de production automatique de mots-clés. Ainsi, les résultats reportés dans la littérature ne sont souvent pas comparables, et il est difficile d'évaluer dans son ensemble les progrès accomplis dans le domaine. Un exemple illustre parfaitement ce problème : il n'est pas possible de contraster les résultats de (Meng et al., 2017), (Florescu and Caragea, 2017) et (Teneva and Cheng, 2017), trois études sur la production automatique de mots-clés publiées la même année à la conférence ACL, car ni les ensembles de données, ni les mesures d'évaluation, ni les pré-traitements linguistiques n'y

sont comparables. En réponse à ce problème, nous avons réalisé une évaluation empirique à grande échelle des modèles de production automatique de mots-clés de l'état-de-l'art (Gallina et al., 2020), dont les résultats, synthétisés ci-après, nous ont permis d'apporter un éclairage nouveau sur leurs forces et leurs faiblesses.

La Table 3.5 présente une partie des résultats obtenus par un éventail de modèles sur neuf ensembles de données de natures différentes (i. e. articles scientifiques, notices bibliographiques et articles de presse) et couvrant plusieurs types d'annotation en mots-clés (i. e. renseignés par les auteurs, des lecteurs ou des indexeurs professionnels). La mesure d'évaluation reportée est la moyenne des précisions moyennes (mAP), calculée par correspondance stricte entre les mots-clés produits et ceux de référence. Pour une comparabilité directe des résultats, tous les modèles présentés s'appuient sur une chaîne de traitement unifiée. Les modèles neuronaux, qui nécessitent de grandes quantités de données d'entraînement, sont appris sur les ensembles KP20k (Meng et al., 2017) pour les textes scientifiques et KPTimes (Gallina et al., 2019) pour les articles de presse. Les autres modèles supervisés sont entraînés soit sur les sous-ensembles d'entraînement lorsqu'ils existent, soit par validation croisée d'un contre tous.

		Textes pleins			*Notices bibliographiques*			*Articles de presse*		
	Modèle	**Pu**^{A}	**Ac**^{A}	**Se**$^{A\cup L}$	**In**^{I}	**Ww**^{A}	**Kp**^{A}	**Du**^{L}	**Cr**^{L}	**Ti**^{I}
①	FirstPhrases	14.7	13.5	10.5	27.9	9.8	12.6	22.3	16.5	8.4
	TextRank	1.8	2.4	2.3	31.4	5.6	7.4	19.4	9.5	2.5
	TF×IDF	16.9	11.4	12.7	**34.4**	10.1	12.3	21.6	15.8	9.4
②	PositionRank	4.6	4.9	4.1	32.2	11.6	11.2	**28.0**	12.7	6.6
	MultipartiteRank	15.0	11.0	10.6	29.0	10.4	13.3	24.9	**17.0**	10.1
	EmbedRank	3.2	2.1	2.0	32.5	10.7	10.0	27.5	12.4	3.3
③	Kea	18.6	13.3	**14.7**	33.2	10.9	13.8	24.5	16.7	10.8
	CopyRNN	**25.4**	**26.3**	13.8	26.4	**24.9**	**28.7**	7.2	4.2	**50.9**
	CorrRNN	19.4	20.5	10.9	23.6	20.3	22.7	6.5	3.2	20.3

Table 3.5. – Résultats en matière de mAP des différents modèles sur les ensembles de données PubMed (Schutz, 2008), ACM (Krapivin et al., 2009), SemEval (Kim et al., 2010), Inspec (Hulth, 2003), WWW (Caragea et al., 2014), KP20k (Meng et al., 2017), DUC-2001 (Wan and Xiao, 2008b), 500N-KPCrowd (Marujo et al., 2012) et KPTimes (Gallina et al., 2019). Pour chaque ensemble, le type d'annotation des mots-clés de référence (i. e. Auteur, Lecteur ou Indexeur) est indiqué en exposant. Les scores soulignés indiquent une amélioration significative par rapport aux *baselines* (t.test < 0.05).

Dans une première série de résultats (voir ① dans la Table 3.5), nous comparons les scores de trois *baselines* : FirstPhrases, qui extrait les N premiers mot-clés candidats du document ; TextRank, une méthode de graphes présentée en détails en §2 ; et TF×IDF, qui extrait les mot-clés candidats de plus hauts poids selon le schéma de pondération éponyme. Sans surprise, TF×IDF, qui contrairement aux

deux autres *baselines* utilise des statistiques calculées sur la collection, donne en général les meilleurs résultats. FirstPhrases, qui s'appuie sur le fait que les idées les plus importantes sont introduites en premier dans les textes (Marcu, 1997), obtient des scores corrects dans l'ensemble, et même parfois meilleurs que ceux des autres *baselines* sur les articles de presse de part leur structuration particulière (i. e. attaque, corps et chute).

Nous comparons dans la seconde série de résultats (voir (2) dans la Table 3.5) trois modèles d'extraction de mots-clés non supervisés : PositionRank et MultipartiteRank, deux méthodes de graphes décrites en §2.2 ; et EmbedRank (Bennani-Smires et al., 2018), qui exploite des représentations vectorielles denses pour ordonner les mots-clés candidats par rapport à leur distance vis-à-vis du document. Bien que MultipartiteRank obtienne les meilleurs résultats de la série, on note des scores du même acabit que ceux des *baselines* sur les textes pleins. Ce constat révèle la difficulté qu'ont les modèles non supervisés à traiter les documents longs, qui s'explique en grande partie par la taille substantiellement plus grande de l'ensemble de mots-clés candidats à ordonner.

La dernière série de résultats (voir (3) dans la Table 3.5) regroupe trois modèles supervisés : Kea (Witten et al., 1999), qui utilise un algorithme de classification naïve bayésienne sur deux traits caractéristiques (i. e. première position et poids TF×IDF) ; ainsi que CopyRNN et CorrRNN, deux méthodes neuronales de génération de mots-clés de l'état de l'art décrites en §3.2.1. On relève la supériorité des modèles supervisés lorsque ceux-ci sont entraînés à partir de données de même nature (i. e. même domaine, même type d'annotations) mais surtout une chute brutale des performances lorsque ce n'est pas le cas. Les scores successivement très élevés puis très faibles sur les ensembles d'articles de presse illustrent le problème de généralisation dont souffrent particulièrement les méthodes neuronales. Malgré sa simplicité et sa relative ancienneté, Kea offre un niveau de performance globalement supérieur à celui des modèles non neuronaux, et constitue une *baseline* que l'on peut toujours qualifier de forte.

Les résultats reportés dans la Table 3.5 mettent en évidence des différences importantes entre les scores des modèles selon le type d'annotation des mots-clés de référence. Ainsi, les performances maximales sont bien plus élevées sur les ensembles de données annotés par des indexeurs professionnels que sur les autres, sans doute en raison de l'homogénéité supérieure de ce type d'annotation. Pour quantifier plus précisément ces différences, nous comparons dans la Table 3.6 les résultats obtenus par chacun des modèles sur une sous-partie de l'ensemble de données Inspec pour lequel nous disposons de deux jeux de mots-clés de référence, le premier donné par des indexeurs professionnels et le second renseigné par les auteurs. On relève une diminution de moitié des scores avec les mots-clés auteurs, ce qui indique une sous-estimation forte des performances des modèles reportées dans les précédentes études. Plus important encore, la supériorité des modèles neuronaux qui se manifeste clairement avec les mots-clés auteurs ne se retrouve pas avec les mots-clés indexeurs. Ce constat souligne à nouveau le problème de généralisation des méthodes neuronales qui n'atteignent leur plein potentiel que sur des données de même nature et type d'annotation que celles utilisées pour l'entraînement.

Modèle	Indexeur	Auteur
FirstPhrases	26.1	13.2 (−49%)
TextRank	29.6	9.3 (−69%)
TF×IDF	**33.3**	16.1 (−52%)
PositionRank	31.0	13.0 (−58%)
MultipartiteRank	27.6	13.6 (−51%)
EmbedRank	31.3	11.5 (−63%)
Kea	31.9	15.9 (−50%)
CopyRNN	29.8	**33.8** (+13%)
CorrRNN	24.2	28.2 (+17%)

Table 3.6. – Résultats en matière de mAP des différents modèles sur un sous-ensemble de 55 documents de Inspec (Hulth, 2003) en fonction du type d'annotation en mots-clés (i. e. indexeurs professionnels ou auteurs). Les scores soulignés indiquent une amélioration significative (t.test < 0.05).

3.4. Discussion

Dans ce chapitre, nous avons relevé les limites de l'évaluation intrinsèque des méthodes d'indexation par mots-clés, et étudié deux cadres applicatifs pour l'évaluation extrinsèque : la recherche documentaire et le résumé automatique. Plusieurs observations importantes ressortent de nos expériences. Tout d'abord, et cela paraît évident, l'évaluation automatique des mots-clés par rapport à une vérité terrain est peu fiable et devrait idéalement servir tout au plus à la validation des modèles. L'évaluation indirecte des mots-clés au travers d'une tâche, que nous prônons ici, est certainement la solution la plus efficace pour faire face à ce problème, mais à la condition de disposer des données nécessaires à son exécution. Lorsque cela n'est pas possible, une autre solution pour augmenter la fiabilité des conclusions tirées des résultats consiste à multiplier les ensembles de données de test au sein d'un cadre expérimental unifié. [5] Cette solution devrait d'ailleurs être systématiquement retenue compte tenu du nombre grandissant d'ensembles de données de test désormais disponibles.

Ensuite, les méthodes neuronales génératives représentent sans conteste l'état de l'art dans la production automatique de mots-clés. Il convient cependant de modérer ce constat par deux remarques liées : premièrement, ces méthodes réclament de grandes quantités de données d'entraînement ; deuxièmement, les modèles de séquence-à-séquence sur lesquels sont construits ces méthodes généralisent mal sur des données aux propriétés différentes de celles de l'entraînement. Dit autrement, l'entraînement de méthodes génératives n'est possible que pour un nombre restreint de domaines (et langues), et leur

5. À cet effet, nous avons émis un ensemble de recommandations (i. e. quels sont les ensembles de données, les mesures d'évaluation et les *baselines* à utiliser) visant à renforcer le protocole expérimental des travaux sur la production automatique de mots-clés (Gallina et al., 2020).

application sur des données différentes engendre une chute de performance conséquente. Des travaux misent sur l'apprentissage auto-supervisé (*self-supervised learning*) pour répondre à ce problème, mais les résultats sont encore très en deçà de ceux des méthodes supervisées (Shen et al., 2021; Wu et al., 2022a). Les travaux sur l'adaptation au domaine des modèles séquence-à-séquence représentent également une piste à considérer (Gururangan et al., 2020; Diao et al., 2021), mais elle n'a pour le moment pas fait l'objet de recherches particulières dans le cadre de la production automatique de mots-clés.

Un autre constat qui découle de nos expériences est que les méthodes neuronales génératives produisent en pratique très peu de mots-clés absents du texte source, e. g. seul 3% des mots-clés produits par la méthode CopyRNN sur l'ensemble de données NTCIR-2 sont absents (Boudin and Gallina, 2021). Pourtant, ce sont ces derniers qui, du fait de leur propriété d'étendre le contenu des documents, ont un impact le plus important sur l'efficacité des systèmes de recherche d'information. Quelques travaux se sont attaqués à cet enjeu, notamment en explorant de manière empirique la paramétrisation des modèles (Meng et al., 2021; Ye et al., 2021a), mais le problème reste entier car il est foncièrement lié à notre capacité pour le moment très limitée d'évaluer avec fiabilité la qualité des mots-clés absents générés.

All have their worth and each contributes to the worth of the others.

J.R.R. Tolkien, The Silmarillion

4

Applications de la recherche d'information

Sommaire

Ce chapitre synthétise les travaux que nous avons réalisés dans différents cadres applicatifs de la recherche d'information, avec pour trame de fond l'association de techniques et outils issus du TAL. Nous commençons par décrire une série de travaux sur la recherche d'information dans le domaine de la médecine clinique (§4.1), puis nous présentons une approche de contextualisation de tweets construite sur le principe d'un modèle de résumé automatique multi-document (§4.2).

4.1. Recherche d'information clinique

Le volume de littérature scientifique dans le domaine de la santé et du biomédical a augmenté de façon exponentielle ces dernières années. Plus d'un million et demi de publications scientifiques ont été indexées sur PubMed en 2020, soit près de trois publications chaque minute.[1] Rechercher une information spécifique dans une telle quantité de données est une tâche difficile que les médecins et les chercheurs sont souvent incapables d'exécuter dans un temps raisonnable. Faciliter l'accès aux publications scientifiques, et plus spécifiquement aux études cliniques est pourtant critique car cela a des répercussions directes sur la prise en charge des patients.

1. https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

La médecine fondée sur les faits (*Evidence-Based Medicine*) (Sackett, 1997) est une pratique largement adoptée en médecine clinique qui consiste à utiliser les meilleures données disponibles pour la prise de décisions concernant les soins à prodiguer à chaque patient. Dans cette pratique, les médecins sont formés à exprimer leur besoin d'information sous la forme de questions structurées autour de quatre éléments désignés par l'abréviation PICO (Richardson et al., 1995; Schardt et al., 2007) : *Population-Problem* (population-problème), *Intervention* (intervention ou stratégie de prise en charge), *Comparison* (comparateur) et *Outcome* (critère de jugement). Par exemple, dans la question clinique « *In children with pain and fever, how does paracetamol compared with ibuprofen affect levels of pain and fever ?* » les quatre éléments PICO ci-dessous sont présents :

Population-Problem	children with pain and fever
Intervention	paracetamol
Comparison	ibuprofen
Outcome	levels of pain and fever

On retrouve cette structuration en éléments PICO dans la plupart des notices bibliographiques d'articles scientifiques médicaux présents dans PubMed, mais rarement mise en évidence de manière explicite (Dawes et al., 2007). Les travaux présentés ci-après explorent l'utilisation des éléments PICO, à la fois dans les requêtes et les documents, pour améliorer l'efficacité des systèmes de recherche d'information dans le domaine médical.

4.1.1. Indexation et recherche par éléments PICO

L'approche de recherche d'information la plus directe consiste à faire correspondre les éléments PICO de la requête avec ceux des documents. Pour cela, il est nécessaire d'identifier les unités textuelles (i.e. mots, expressions polylexicales) correspondant à chacun des éléments dans les documents. Or, ce type d'annotation n'est pas possible à ce niveau de détails, notamment en raison de désaccords non résolubles entre annotateurs sur la délimitation des éléments (Demner-Fushman et al., 2006). À l'instar des travaux précédents (Demner-Fushman and Lin, 2007; Chung, 2009), nous levons cette difficulté par l'identification des éléments PICO à un niveau de granularité plus élevé, celui de la phrase. Pour ce faire, nous avons entraîné un modèle ensembliste de classification de phrases sur un ensemble de données annotées semi-automatiquement à l'aide de patrons linguistiques (Boudin et al., 2010a).

La Table 4.1 présente les résultats obtenus par notre modèle ensembliste pour chacun des éléments PICO. La mesure d'évaluation reportée est la f-mesure calculée par validation croisée sur 10 partitions. Cinq algorithmes de classification sont combinés [2] pour prédire l'élément PICO d'une phrase à partir

2. Nous utilisons l'implémentation des algorithmes de classification de l'outil Weka (Hall et al., 2009) avec les paramètres par défaut. Le modèle ensembliste consiste en une combinaison linéaire des scores de prédiction calculés par chaque algorithme de classification.

d'un ensemble de traits statistiques (e. g. longueur, position) et calculés à l'aide de ressources externes (e. g. thésaurus MeSH [3]). Comme attendu, le modèle ensembliste améliore les performances par rapport aux algorithmes de classification pris individuellement. On note la grande précision avec laquelle le modèle est capable d'identifier les phrases correspondant à l'élément P, mais aussi les scores nettement plus faibles pour les autres éléments. Deux raisons expliquent ces résultats : la quantité décroissante de données d'entraînement (i. e. 14 279 documents pour l'élément P, 9 095 pour l'élément I/C et 2 394 pour l'élément O) ; et la difficulté intrinsèque de la tâche (e. g. il existe souvent plusieurs résultats (élément O) dans une étude clinique).

Modèle	**P**	**I/C**	**O**
Ensemble	86.3	67.0	56.6
⊢ Modèle 1 (J48)	77.7	55.9	45.5
⊢ Modèle 2 (NaiveBayes)	66.0	49.0	48.1
⊢ Modèle 3 (RandomForest)	83.9	63.5	50.6
⊢ Modèle 4 (SVM)	74.3	39.3	19.0
⊢ Modèle 5 (MultilayerPerceptron)	85.4	66.3	55.7

Table 4.1. – Résultats en matière de f-mesure (F_1) des différents modèles de classification de phrases en fonction de l'élément PICO à identifier.

Dans une première série d'expériences, nous avons évalué l'approche de recherche d'information qui consiste à faire correspondre les éléments PICO identifiés dans les documents avec ceux de la requête (Boudin et al., 2010e). Les résultats négatifs que nous avons obtenus pointent le caractère trop restrictif de cette approche, qui est exacerbé par la faible précision de l'identification automatique des éléments PICO. Pourtant, même si elle est perfectible, cette annotation des éléments PICO pourrait être suffisante pour pondérer plus précisément l'importance des termes dans les documents. Pour le vérifier, trois extensions du système de recherche d'information par modèle de langue (Ponte and Croft, 1998) sont proposées et comparées.

De façon formelle, le score de pertinence d'un document d par rapport à une requête q est déterminé par le calcul de la divergence de Kullback-Leibler entre leurs modèles de langue respectifs (Zhai and Lafferty, 2001) selon l'équation (4.1) donnée ci-dessous.

$$\begin{aligned} score(q,d) &= -D_{\mathrm{KL}}(Q||D) \\ &\propto -\sum_{w \in q} p(w|\hat{\theta}_q) \cdot \log p(w|\hat{\theta}_d) \end{aligned} \tag{4.1}$$

3. https://www.nlm.nih.gov/mesh/

où $\hat{\theta}_d$ et $\hat{\theta}_q$ sont les modèles de langue uni-gramme du document d et de la requête q, respectivement. La probabilité $p(w|\hat{\theta}_d)$ est estimée par maximum de vraisemblance avec lissage de Dirichlet (Zhai and Lafferty, 2004) selon l'équation (4.2) donnée ci-dessous.

$$p(w|\hat{\theta}_d) = \frac{\mathrm{c}(w,d) + \mu \cdot p(w|C)}{|d| + \mu} \tag{4.2}$$

où $\mathrm{c}(w,d)$ est la fréquence du mot w dans le document d, $|d|$ la taille du document d et $p(w|C) = \frac{c(w,C)}{|C|}$ la probabilité du mot w dans la collection C.

La première extension (ω_d) consiste en une surpondération des mots du document en fonction de l'élément PICO dans lequel ils apparaissent. Pour cela, un modèle de langue $\hat{\theta}_e$ est construit pour chaque élément e et la probabilité $p(w|\hat{\theta}_d)$ est redéfinie selon l'équation (4.3) ci dessous.

$$p'(w|\hat{\theta}_d) = p(w|\hat{\theta}_d) + \sum_{e \in \{P,I,C,O\}} \alpha_e \cdot p(w|\hat{\theta}_e) \tag{4.3}$$

La seconde extension (ω_q) consiste en une surpondération des mots de la requête en fonction de l'élément PICO auquel ils appartiennent. Soit q_e le sous-ensemble de mots de la requête q appartenant à l'élément e, le score de pertinence d'un document d par rapport à une requête q est obtenu selon l'équation (4.4) ci dessous.

$$score'(q,d) = score(q,d) + \sum_{e \in \{P,I,C,O\}} \beta_e \cdot score(q_e, d) \tag{4.4}$$

La troisième extension ($\omega_{d,q}$) est simplement l'application conjointe de la surpondération des éléments PICO du document et de la requête.

La Table 4.2 présente les résultats obtenus par le système de recherche d'information par modèle de langue et les trois extensions proposées sur une collection de test composée de 52 requêtes formulées par des médecins et annotées manuellement en éléments PICO.[4] Il en ressort que l'annotation des éléments PICO permet d'améliorer significativement les scores, mais que cette amélioration est principalement attribuable à la surpondération des éléments de la requête. La surpondération des éléments des documents offre bien un gain de performance supplémentaire, confortant l'idée que cette information est utile pour la recherche d'information, mais sa portée reste limitée à cause notamment de la faible précision de l'identification automatique des éléments PICO. Nous allons voir dans la section suivante une réponse à ce problème construite sur les distributions des éléments PICO dans les documents.

4. Nous utilisons l'implémentation de l'outil `indri` (Strohman et al., 2005) avec les paramètres par défaut. Les paramètres α et β introduits dans les extensions sont déterminés par recherche en grille en validation croisée sur 2 partitions.

Métrique \ Modèle	KL	$+\omega_d$	$+\omega_q$	$+\omega_{d,q}$
mAP	11.6	11.7 (+0.5%)	<u>15.1</u> (+30.2%)	<u>15.2</u> (+30.9%)
P@10	25.0	25.0 (+0.0%)	<u>35.4</u> (+42.7%)	<u>35.8</u> (+43.0%)

Table 4.2. – Résultats en matière de P@10 et de mAP du système de recherche d'information par modèle de langue et des trois extensions de pondération des éléments PICO. Les scores soulignés indiquent une amélioration significative (t.test < 0.01).

4.1.2. Distribution des éléments PICO dans les documents

Les éléments PICO ne sont pas uniformément distribués dans les documents mais se répartissent de manière spécifique dans la structuration standardisée généralement utilisée dans le domaine médical, e. g. le format IMRAD (Introduction, Methods, Results, Discussion) (Sollaci and Pereira, 2004). Une alternative à la surpondération des éléments PICO consisterait donc à surpondérer les passages dont la probabilité de contenir ces éléments est élevée (Boudin et al., 2010b). Pour vérifier le bien fondé de cette hypothèse, nous avons calculé, pour chaque élément, la distribution des mots utilisés dans les requêtes au sein des documents pertinents. La Figure 4.1 montre une distribution caractéristique des éléments PICO « en U », qui indique une surreprésentation de ces derniers dans les parties introductives et conclusives des documents.

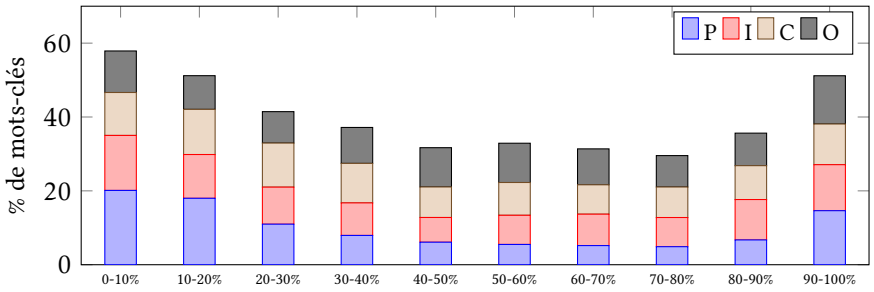


Figure 4.1. – Distribution, pour chaque élément PICO, des mots de la requête dans les différentes parties des documents pertinents sur la collection de test CLIREC (Boudin et al., 2010c).

Nous proposons en conséquence une nouvelle extension (w_{plm}) du système de recherche d'information par modèle de langue (équation (4.1)) qui consiste en une surpondération des mots du document en fonction de leur position (Boudin et al., 2010d). Pour cela, un modèle de langue $\hat{\theta}_p$ est construit pour chaque partie p du document découpé en $n = 10$ parties égales en nombre de mots, et la probabilité $p(w|\hat{\theta}_d)$ est redéfinie selon l'équation (4.5) ci dessous.

$$p''(w|\hat{\theta}_d) = \gamma_1 \cdot p(w|\hat{\theta}_d) + \gamma_2 \cdot p(w|\hat{\theta}_{titre}) + \gamma_3 \cdot \sum_{p \in \{p0,\ldots,p9\}} \delta_{p,e} \cdot p(w|\hat{\theta}_p) \tag{4.5}$$

où les paramètres $\delta_{p,e}$, qui déterminent l'importance de l'élément e dans la partie p du document, sont fixés aux valeurs des distributions des éléments PICO observées dans les documents.

La surpondération des éléments PICO de la requête (w_q, équation (4.4)) est également appliquée, mais sous une forme simplifiée. Soit q_e le sous-ensemble de mots de la requête q appartenant à l'élément e, le score de pertinence d'un document d par rapport à une requête q est obtenu selon l'équation (4.6) ci dessous.

$$score''(q,d) = \sum_{e \in \{P,I,C,O\}} \beta_e \cdot score(q_e, d) \tag{4.6}$$

La Table 4.3 présente les résultats obtenus par le système de recherche d'information par modèle de langue et des extensions proposées sur la collection de test CLIREC (Boudin et al., 2010c) (présentée en détails en §5.1.2). [5] Les métriques d'évaluation reportées sont la mAP, la P@5 et le nombre de documents pertinents retrouvés. On note un gain significatif de performance lorsque la surpondération positionnelle des mots est appliquée, qui vient s'ajouter à l'amélioration déjà apportée par la surpondération des éléments PICO de la requête. Ce résultat confirme la pertinence du critère de position pour la surpondération des éléments marquants dans les documents lorsque la précision de leur identification automatique n'est pas satisfaisante. De part sa généricité, notre méthode peut également être étendue aux systèmes de recherche d'information exploitant la structure des documents, i. e. qui pondèrent d'avantage les mots apparaissant dans certaines parties (champs) du documents (Robertson et al., 2004; Kim and Croft, 2012).

Métrique \ **Modèle**	KL	$+w_q$	$+w_{q,plm}$
mAP	12.6	14.4 (+14.3%)	<u>16.3</u> (+29.4%)
P@5	17.2	<u>19.6</u> (+14.0%)	<u>24.0</u> (+39.5%)
#docs	5433	5780 (+6.4%)	5770 (+6.2%)

Table 4.3. – Résultats en matière de mAP, de P@5 et de nombre de documents pertinents retrouvés (#docs) du système de recherche d'information par modèle de langue et des extensions sur la collection de test CLIREC. Les scores soulignés indiquent une amélioration significative (t.test < 0.01).

4.2. Contextualisation de tweets

Twitter est un réseau social de microblogage qui permet à des particuliers et à des entreprises de diffuser des messages courts (280 caractères maximum) appelés tweets, immédiatement visibles par

5. Les paramètres β et γ introduits dans les extensions sont déterminés en optimisant la métrique mAP par recherche en grille en validation croisée sur 10 partitions.

les autres utilisateurs du monde entier. Les utilisateurs de ce réseau s'expriment donc succinctement, et la teneur des messages peut parfois être, faute de contexte, difficile à appréhender (Bellot et al., 2016). La tâche de contextualisation de tweets vise à répondre à ce problème par la génération d'un résumé explicitant le contexte d'un tweet pour en faciliter la compréhension. La suite de cette section décrit plus en détails cette tâche et présente une partie les résultats que nous avons obtenus avec une approche mêlant recherche d'information et résumé automatique multi-document (Deveaud and Boudin, 2012, 2013b,a, 2014).

4.2.1. Description de la tâche

Les travaux que nous présentons ici portent sur la tâche de contextualisation de tweets, telle que définie au sein des campagnes d'évaluation INEX organisées entre 2012 et 2014. [6] L'objectif de cette tâche est d'aider un utilisateur à mieux comprendre un tweet en lui fournissant un court résumé explicatif (500 mots maximum) produit automatiquement à partir de passages extraits de l'encyclopédie libre Wikipédia (voir la Figure 4.2). Pour évaluer et comparer les approches de contextualisation de tweets, nous disposons du cadre expérimental établi par les organisateurs de la campagne INEX qui se compose d'une collection de test et de métriques spécifiques.

tweet : Bobby Brown – Fighting #WhitneyHouston's Family to See Bobbi Kristina

résumé : Whitney Elizabeth Houston (August 9, 1963 – February 11, 2012) was an American recording artist, actress, producer, and model. Houston was one of the world's best-selling music artists, having sold over 170 million albums, singles and videos worldwide. Robert Barisford "Bobby" Brown (born February 5, 1969) is an American R&B singer-songwriter, occasional rapper, and dancer. After a three-year courtship, the two were married on July 18, 1992. On March 4, 1993, Houston gave birth to their daughter Bobbi Kristina Houston Brown, her only child, and his fourth. With the missed performances and weight loss, rumors about Houston using drugs with her husband circulated. Following fourteen years of marriage, Brown and Houston filed for legal separation in September 2006. Their divorce was finalized on April 24, 2007, with Houston receiving custody of their then-14-year-old daughter. On February 11, 2012, Houston was found unresponsive in suite 434 at the Beverly Hilton Hotel, submerged in the bathtub

Figure 4.2. – Exemple de résumé contextuel repris de (Bellot et al., 2016) dans lequel toutes les phrases proviennent de pages Wipipédia différentes.

4.2.2. Approche proposée

L'approche que nous proposons met en jeu successivement un système de recherche d'information pour retrouver les articles Wikipédia en rapport avec un tweet, et un système de résumé automatique

6. https://inex.mmci.uni-saarland.de/

multi-document pour former un contexte explicatif à partir des passages les plus pertinents (Deveaud and Boudin, 2014, voir Figure 4.3). Plus spécifiquement, nous employons une extension (*Markov Random Field*, abrégé ci-après MRF) du système de recherche d'information par modèle de langue qui tient compte des dépendances entre les termes de la requête (Metzler and Croft, 2005). Les tweets, utilisés ici comme des requêtes, sont tout d'abord nettoyés et les *hashtags* qu'ils contiennent sont repérés et découpés automatiquement (e.g. `#didyouknow` → `did you know`) à l'aide d'un modèle de langue estimé sur le corpus Bing Web N-gram (Gao et al., 2010). De façon formelle, le score de pertinence d'un document d par rapport à un tweet t est déterminé selon l'équation (4.7) donnée ci-dessous.

$$score(t,d) = \alpha \cdot score_{\text{MRF}}(t,d) + (1-\alpha) \cdot score_{\text{MRF}}\Big(ht(t),d\Big) \tag{4.7}$$

où $ht(t)$ est l'ensemble des *hashtags* contenus dans le tweet t, et α permet de calibrer l'importance donnée aux hashtags par rapport au reste du tweet. [7]

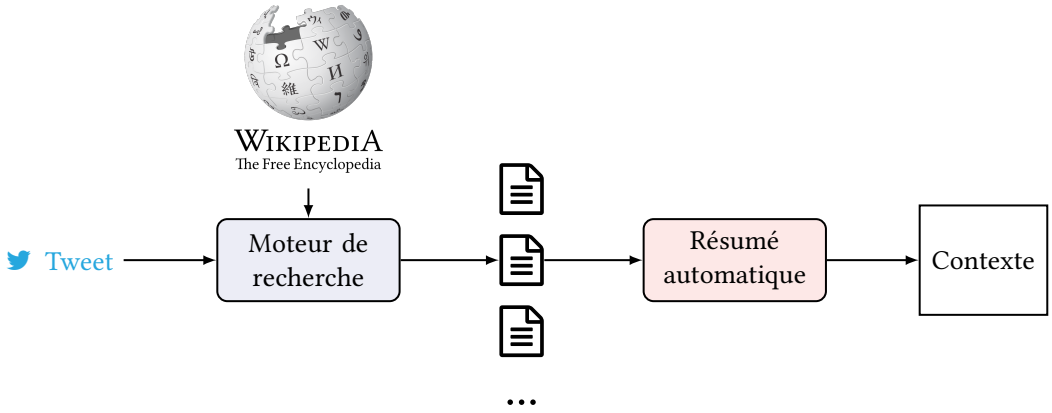


Figure 4.3. – Illustration de l'approche de contextualisation de tweets à partir de Wikipédia.

Le contexte explicatif d'un tweet est obtenu par l'assemblage de phrases extraites des n documents les plus pertinents. Chaque phrase reçoit un score d'importance calculé à partir de trois critères : son importance vis-à-vis du document d'où elle provient (estimée avec TextRank, voir §2), sa pertinence par rapport au tweet (mesurée par similarité lexicale) et la pertinence du document d'où elle provient. Un algorithme glouton est ensuite utilisé pour sélectionner l'ensemble optimal de phrases pour le contexte, i. e. ne dépassant pas 500 mots et dont le score global (importance et diversité) est le plus élevé.

La Table 4.4 présente une partie des résultats officiels de la campagne INEX 2013 de contextualisation de tweets à laquelle nous avons participé avec l'approche décrite ci-dessus. A l'instar du résumé automatique, les contextes sont évalués en fonction de deux critères : l'informativité et la lisibilité.

7. Nous utilisons les paramètres recommandés par les auteurs pour le modèle MRF, à savoir $\lambda_T = 0.85$, $\lambda_O = 0.10$ et $\lambda_U = 0.05$.

Pour le premier critère, la métrique reportée est la divergence (Δ) entre les contextes générés et les contextes de référence, calculée sur les bi-grammes à trou de mots racinisés. Pour le second critère, des évaluations manuelles ont été réalisées pour mesurer la pertinence, la redondance, la validité et la syntaxe des contextes. Notre approche (199/258) obtient les meilleurs résultats de la campagne selon la métrique officielle (Δ), et se classe entre la seconde et la troisième position au niveau de la lisibilité sans que nous n'ayons mis oeuvre de techniques pour optimiser cet aspect. L'étude des critères pour le calcul de l'importance des phrases montre que le score TextRank et la proximité lexicale avec le tweet sont les plus utiles.

Métrique / Part./Run	Δ	Pert.	Red.	Val.	Syn.
199/258	**0.8943**	68.36%	64.52%	66.04%	67.34%
182/275	0.8969	76.64%	67.30%	74.52%	75.50%
65/254	0.9242	73.30%	61.52%	68.94%	71.92%
62/276	0.9301	52.08%	45.84%	51.24%	52.08%
46/270	0.9397	46.84%	41.20%	45.30%	46.00%

Table 4.4. – Résultats officiels en matière d'informativité (Δ), de pertinence, de redondance, de validité et de syntaxe du meilleur *run* pour les 5 meilleurs participants à la campagne INEX 2013 de contextualisation de tweets.

4.3. Discussion

Dans ce chapitre, nous avons présenté deux séries de travaux de recherche mêlant fructueusement recherche d'information et techniques du TAL. Les résultats obtenus dans le domaine médical (§4.1) permettent de nuancer le postulat sur l'inutilité des techniques du TAL pour la recherche d'information (Jones, 1999), tandis que ceux obtenus sur la contextualisation de tweets (§4.2) mettent en évidence les bénéfices de l'utilisation de moteurs de recherche pour une tâche du TAL. D'autres travaux que nous avons menés, notamment sur l'utilisation de thésaurus pour la prédiction de la difficulté des requêtes (Boudin et al., 2012b) et sur la correction automatique d'erreurs de reconnaissance optique de caractères (OCR) pour la recherche de livres (Deveaud et al., 2011b,a), viennent appuyer ce constat d'une interaction mutuellement profitable entre TAL et recherche d'information.

Nos observations font écho à celles rapportées par Claveau (2020) sur un rapprochement *inexorable* entre les communautés TAL et RI. Dans les faits, il aura fallu attendre l'arrivée des modèles de langue pré-entraînés construits autour de l'architecture *transformers* (Vaswani et al., 2017), et plus précisément de BERT (Devlin et al., 2019), pour voir ce rapprochement véritablement effectif (Lin, 2021). Des différences persistent malgré-tout dans l'application de ces modèles au sein des deux communautés,

et en particulier dans leurs objectifs de correspondance sémantique (*semantic matching*) pour le TAL et correspondance de pertinence (*relevance matching*) en RI (Rao et al., 2019). D'autres enjeux viennent accentuer ces différences comme le maintien d'un temps de latence acceptable pour la RI qui s'oppose directement au temps de calcul prohibitif introduit par l'utilisation de modèles de langue pré-entraînés (Lin, 2022).

May the hair on your toes never fall out!

J.R.R. Tolkien, The Hobbit

5 Ressources, outils et valorisation

Sommaire

Les ressources langagières et les outils logiciels sont au cœur des travaux de recherche menés en TAL et en RI. La prépondérance des méthodes statistiques entraîne un besoin impérieux de données, qui ne cesse de prendre de l'ampleur avec la transition de la communauté scientifique vers les architectures neuronales profondes. Le partage du code source, et plus récemment des modèles pré-entraînés, est également un enjeu majeur puisqu'il conditionne largement la reproductibilité des résultats (Wieling et al., 2018; Cohen et al., 2018). Ce chapitre condense les travaux que nous avons réalisés dans la construction de ressources (§5.1), le développement d'outils (§5.2) et leur valorisation dans la communauté scientifique (§5.3).

5.1. Ressources langagières

Cette section présente les ensembles de données que nous avons produit en les regroupant selon l'application visée.

5.1.1. Indexation par mots-clés

Les ensembles de données pour l'indexation par mots-clés sont constitués d'une collection de documents associés à des mots-clés de référence. Ces derniers peuvent être annotés par les auteurs des documents eux-mêmes ou par des tiers (lecteurs ou indexeurs professionnels). Plusieurs éléments

agissent sur la qualité générale et l'utilisabilité des ensembles de données, e. g. le type d'annotation, la nature et la couverture thématique des documents, la quantité de documents annotés. Au final, peu de travaux se sont véritablement penchés sur la création de données pour l'indexation par mots-clés, et la plupart de celles disponibles concernent le triptyque « *notices bibliographiques*, *mots-clés auteurs*, *langue anglaise* ». Nous décrivons ci-après les efforts que nous avons entrepris pour combler ce déficit, d'abord sous la forme d'un résumé pour chaque ensemble que nous avons créé, puis par une analyse statistique contrastive. À noter que les ensembles de données présentés sont tous disponibles sur la plateforme Hugging Face.

`wikinews-fr-100` (Bougouin et al., 2013) est un ensemble de données composé de 100 articles de presse en français issus de Wikinews. Chaque document a été annoté en mots-clés par trois lecteurs (étudiants) dans le cadre d'un projet pour le cours Corpus et méthode expérimentale du Master informatique de Nantes Université.

`taln-archives` (Boudin, 2013b) est un ensemble de données composé de 1 207 notices bibliographiques d'articles scientifiques publiés dans les conférences francophones TALN et RECITAL de 1997 à 2015. L'annotation en mots-clés provient des auteurs et une sous-partie des documents (38%) dispose de traductions en anglais. Cet ensemble est dérivé de l'archive numérique francophone des articles de recherche en TAL de l'ATALA.[1]

`termith-eval` (Bougouin et al., 2016a) est un ensemble de données composé de 399 notices bibliographiques en français issues des bases de données PASCAL (sciences exactes) et FRANCIS (sciences humaines) de l'Institut de l'Information Scientifique et Technique (INIST). Les documents couvrent quatre domaines (linguistique, sciences de l'information, archéologie et chimie) et ont été annotés par des indexeurs professionnels dans le cadre du projet ANR TermITH.[2]

`semeval-2010-pre` (Boudin et al., 2016) est un ensemble de données composé de 244 articles scientifiques en anglais issus de la bibliothèque numérique ACM Digital Library. Il s'agit d'une version enrichie et nettoyée des données produites pour la tâche d'extraction automatique de mots-clés de la campagne d'évaluation SemEval-2010 (Kim et al., 2010). Les documents ont été annotés en mots-clés de manière combinée par les auteurs et par des lecteurs (étudiants).

kptimes (Gallina et al., 2019) est un ensemble de données composé de 289 923 articles de presse en anglais collectés sur les sites web du New York Times et du Japan Times. Les documents ont été annotés en mots-clés par des éditeurs de manière semi-automatique : un système d'indexation contrôlée propose un ensemble de mots-clés que les éditeurs peuvent réviser, i. e. ajouter ou supprimer des mots-clés. Des informations additionnelles comme la catégorie thématique (e. g. sports, science, politique) ou l'auteur de l'article ont été extraites des métadonnées des pages web.

1. `http://talnarchives.atala.org/`
2. `https://anr.fr/Project-ANR-12-CORD-0029`

kpbiomed (Houbre et al., 2022) est un ensemble de données composé de 5,6 millions de notices bibliographiques d'articles de recherche biomédicale en anglais issues de PubMed Baseline Repository. Les documents ont été annotés en mots-clés par leurs auteurs et correspondent aux publications de 2011 à 2021. Trois découpages de taille incrémentale pour l'ensemble d'entraînement sont proposés, i. e. petit (500K), moyen (2M) et large (5,6M), et permettent d'étudier l'impact de la quantité de données d'entraînement sur les performances des méthodes.

La Table 5.1 synthétise les caractéristiques des ensembles de données que nous avons créé. Trois ensembles sont en français, et disposent de types d'annotation d'un niveau qualitatif croissant, allant de satisfaisant (annotation par des lecteurs) à très élevé (annotation par des indexeurs professionnels). Ces différences dans l'annotation en mots-clés se retrouvent dans les distributions par catégorie PRMU (voir §3.2) avec une propension marquée des indexeurs à assigner des mots-clés absents. Les trois autres ensembles viennent en réponse aux besoins de la communauté scientifique : `semeval-2010-pre` qui permet de mesurer l'impact des pré-traitements linguistiques sur les performances des modèles de production par mots-clés ; `kptimes` et `kpbiomed` qui permettent d'entraîner des méthodes neuronales de génération de mots-clés dans les domaines journalistique et biomédical respectivement, et d'évaluer leur capacité à généraliser à d'autres domaines.

Ensemble	doc.	mots	nature	lan.	m.-c.	annot.	P R M U
`wikinews-fr-100`	100	307	actualités	fr	9.6	lecteur	
`taln-archives`	1 207	138	notices	fr	4.1	auteur	
`termith-eval`	400	157	notices	fr	11.8	indexeur	
`semeval-2010-pre`	244	8 246	articles	en	15.1	combinée	
`kptimes`	290K	738	actualités	en	5.0	indexeur	
`kpbiomed`	5.6M	271	notices	en	5.2	auteur	

Table 5.1. – Caractéristiques des ensembles de données pour l'indexation par mots-clés que nous avons produits. Nous reportons le nombre de documents (doc.) et de mots, la nature, la langue (lan.), le nombre de mots-clés, le type d'annotation et la distribution par catégorie PRMU des mots-clés.

5.1.2. Recherche d'information

Les ensembles de données en RI sont appelés *collections de test*, et regroupent classiquement une collection de documents et un ensemble de requêtes associées à des jugements de pertinence. Elles sont utilisées pour paramétrer et évaluer l'efficacité des systèmes de RI selon le paradigme dit de *Cranfield* (Cleverdon, 1967), et se retrouvent dans les principaux forums d'évaluation (e. g. TREC, NTCIR, CLEF). La création d'une collection de test est un processus coûteux (Voorhees, 2002), qui nécessite mobilisation d'experts pour la collecte des documents, l'écriture des requêtes et la constitution des

jugements de pertinence associant chaque requête avec un ou plusieurs documents jugés « pertinents ». Nous décrivons ci-après deux initiatives que nous avons menées pour la création de collections de test, avec pour point commun l'extraction semi-automatique de requêtes et de jugements de pertinence à partir du contenu d'articles scientifiques. À noter que les collections de test présentées sont librement disponibles sur la plateforme Github.

`clirec` (Boudin et al., 2010c) est une collection de test pour la RI *ad hoc* dans le domaine scientifique biomédical dérivée de revues systématiques créées par des experts de la fondation Cochrane pour synthétiser les travaux pertinents pour une question clinique (Sackett et al., 1996). Cette collection est composée de 1,2 million de notices bibliographiques issues de PubMed et d'un ensemble de 422 requêtes formulées en éléments PICO (voir §4.1). Les requêtes ont été écrites par des médecins à partir d'une sélection de 155 revues systématiques. Les travaux cités comme valides dans les revues systématiques sont utilisés comme jugements de pertinence et ont été manuellement associés à leurs notices dans la collection.

`acm-cr` (Boudin, 2021) est une collection de test pour la recommandation de citation construite à partir de passages extraits d'articles scientifiques dans le domaine de la RI. La recommandation de citation est la tâche qui consiste à trouver des citations pertinentes pour un *contexte de citation* donné, i. e. une phrase ou un paragraphe dans un document. Cette collection est composée de 114 882 notices bibliographiques d'articles scientifiques issues de l'ACM Digital Library et d'un ensemble de 341 contextes de citation considérés comme requêtes. Les contextes de citation correspondent à des paragraphes entiers, extraits manuellement des textes pleins de 50 articles scientifiques, et pour lesquelles les travaux cités sont utilisés comme jugements de pertinence.

La Table 5.2 synthétise les caractéristiques des collections de test que nous avons construit et donne les résultats obtenus par un système *baseline* de recherche d'information (Okapi-BM25). Comme on pouvait s'y attendre, les performances du système sur la collection `clirec` augmentent nettement avec des requêtes formulées en éléments PICO par rapport aux requêtes par mots-clés (voir ① dans la Table 5.2) Les scores reportés sont en comparaison plus bas ($\approx -10\%$) que ceux obtenus sur d'autres collections de test du domaine biomédical, ce qui s'explique en partie par les biais d'annotation observés sur ces dernières (e. g. *pool* de documents à annoter construit avec Okapi-BM25) (Lipani, 2016, 2019). La seconde série de résultats (voir ② dans la Table 5.2) montre des scores globalement plus élevés pour `acm-cr`, et cette fois-ci meilleurs que ceux obtenus sur d'autres collections de test pour la recommandation de citation (Thakur et al., 2021). Nous attribuons cette différence à deux caractéristiques de la collection `acm-cr` : la qualité des requêtes et des jugements de pertinence extraits manuellement, et l'homogénéité thématique de la collection de documents (i. e. appartenant tous au domaine de recherche d'information).

	Collection	Documents		Requêtes				BM25		
		#nb	mots	type	#nb	mots	qrels	mAP	P@5	nDCG@10
①	clirec	1 211 820	224	m.-c.	155	3.4	2615	12.0	16.1	16.5
				PICO	422	18.3	8887	14.6	21.7	22.8
②	acm-cr	114 882	163	para.	268	145	900	22.4	12.9	28.3
				phr.	552	31	978	28.4	8.4	31.6

Table 5.2. – Caractéristiques des collections de test et résultats obtenus par un système de recherche d'information Okapi-BM25 en matière de mAP, de P@5 et de gain cumulé normalisé (nDCG@10). Le nombre (#nb) de documents/requêtes, le nombre moyen de mots par document/requête, le nombre de documents pertinents (qrels) et le type (mots-clés, PICO, paragraphes ou phrases) des requêtes sont reportés.

5.2. Outils logiciels

Le développement, le maintien et la diffusion d'outils logiciels va de pair avec l'activité de recherche que nous menons en TAL et en RI, nous permettant de tester la validité de nos idées et offrant aux autres chercheurs la possibilité de reproduire, de vérifier ou d'étendre nos travaux. Cette section se concentre sur la bibliothèque logicielle d'extraction de mots-clés `pke`, qui est de loin l'outil logiciel le plus populaire parmi ceux que nous avons développés (e. g. 1.2K+ étoiles, 250 recopies (*forks*) et 13K+ clones sur une période de 14 jours au 03/06/2022). À noter que l'ensemble des outils logiciels développés dans le cadre de nos travaux sont disponibles sous licence libre sur la plateforme GitHub et que nos modèles pré-entraînés sont téléchargeables sur la plateforme Hugging Face.

`pke` (Python Keyphrase Extraction) est une bibliothèque logicielle python sous licence libre dans laquelle sont implémentées les principales méthodes d'extraction de mots-clés de la littérature. Initialement développée dans le cadre du projet CNRS PEPS GOLEM en 2015 puis présentée en démonstration l'année suivante dans la conférence internationale COLING 2016 (Boudin, 2016), `pke` a été progressivement étendue aux méthodes de l'état-de-l'art et sert de socle expérimental au projet ANR JCJC DELICES. À un niveau plus large, `pke` est considérée comme une bibliothèque de référence par la communauté scientifique et a servi pour les sessions de démonstration du tutoriel « *From Fundamentals to Recent Advances: A Tutorial on Keyphrasification* » (Meng et al., 2022) qui s'est déroulé à la conférence internationale ECIR 2022.

De façon plus spécifique, `pke` est organisée en une chaîne de traitement dans laquelle chaque composant peut être modifié ou étendu pour développer de nouvelles méthodes (voir Figure 5.1 à gauche). Les pré-traitements linguistiques (i. e. le découpage du texte en phrases, la *tokenization* et l'étiquetage grammatical) sont externalisés et réalisés à l'aide de la bibliothèque `spacy`, ce qui en

pratique permet d'extraire des mots-clés dans plus d'une vingtaine de langues.[3] La bibliothèque `pke` dispose d'une API minimale, unifiée pour les méthodes supervisées et non-supervisées, qui est illustrée par un exemple dans la Figure 5.1 (à droite). Pour les méthodes supervisées, des modèles entraînés sur l'ensemble de données `semeval-2010-pre` sont distribués et paramétrés par défaut, et des tutoriaux sont mis à disposition pour permettre l'entraînement de nouveaux modèles.

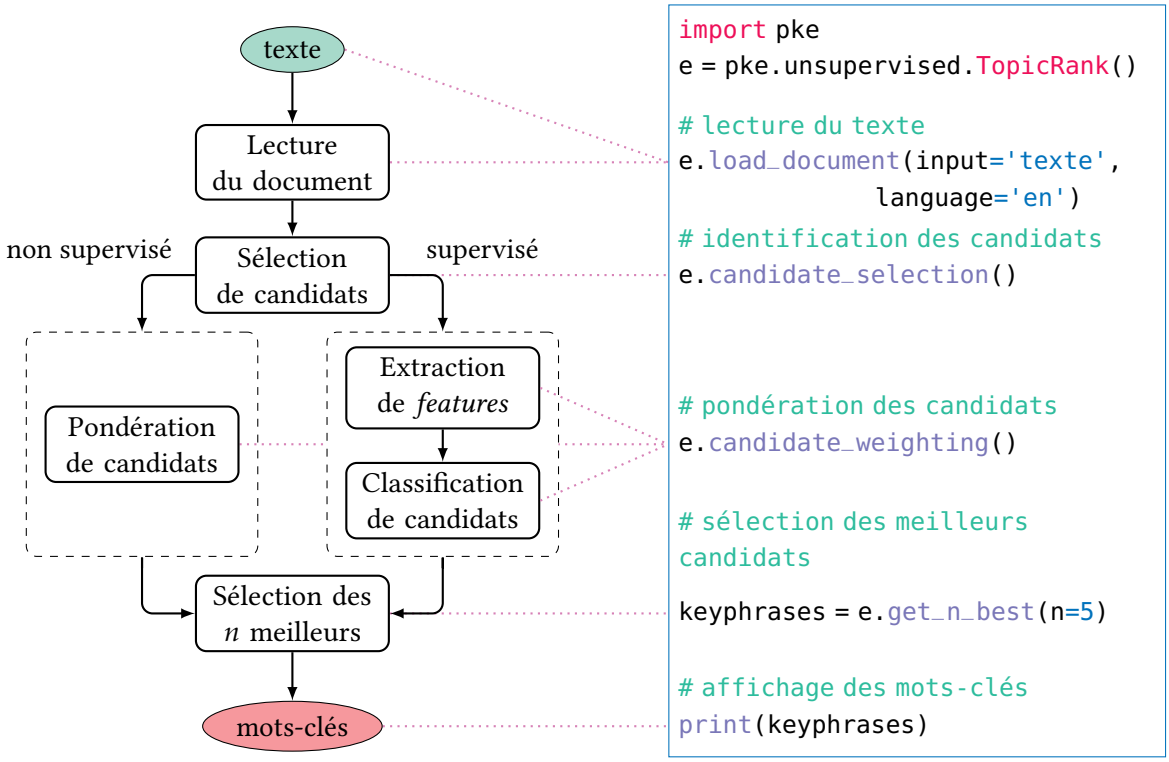


Figure 5.1. – Architecture de la bibliothèque logicielle `pke` (à gauche) associée à un exemple minimal d'utilisation de son API (à droite).

La bibliothèque `pke` fournit actuellement les implémentations de 11 méthodes d'extraction de mots-clés, avec une inclinaison particulière pour les méthodes de graphes non supervisées. Un dépôt complémentaire [4] fournit le code permettant de mettre en comparaison les résultats obtenus par les méthodes implémentées sur les ensembles de données les plus utilisés par la communauté scientifique. Nous y suivons les recommandations émises en §3.3 (i. e. chaîne de traitement unifiée, mesures d'évaluation standardisées) visant à assurer la comparabilité directe des résultats. Un exemple de comparaison des méthodes sur l'ensemble de données `semeval-2010-pre` est montré dans la Table 5.3. Sans entrer dans les détails des performances, il est intéressant de noter la bonne efficacité des implémentations qui pour la plupart permettent le traitement de plusieurs centaines de documents à la seconde (it/s) sur un ordinateur portable ordinaire.

3. https://spacy.io/usage/models#languages
4. https://github.com/boudinfl/pke-benchmarking

		Méthode	it/s	F_1@5	mAP
non supervisé	statistique	FirstPhrases	399	14.0	10.7
		TfIdf	428	12.9	11.1
		KPMiner (El-Beltagy and Rafea, 2010)	427	12.3	10.8
		YAKE (Campos et al., 2020)	13	15.9	11.4
	graphe	TextRank (Mihalcea and Tarau, 2004)	305	9.9	8.3
		SingleRank (Wan and Xiao, 2008a)	256	11.8	9.7
		TopicRank (Bougouin et al., 2013)	200	12.0	8.4
		TopicalPageRank (Sterckx et al., 2015)	21	11.9	10.0
		PositionRank (Florescu and Caragea, 2017)	179	12.8	10.6
		MultipartiteRank (Boudin, 2018)	149	13.9	10.9
sup.		Kea (Witten et al., 1999)	366	14.1	11.6

Table 5.3. – Résultats en matière de nombre de documents traités par seconde (it/s), de F_1@10 et de mAP pour les méthodes implémentées dans `pke` sur l'ensemble de données `semeval-2010-pre` calculés avec un MacBook Air M1.

5.3. Valorisation

Il existe différentes interprétations du terme *valorisation*, aussi nous retenons dans ce manuscrit celle du Comité National d'Évaluation (CNÉ) qui est « *rendre utilisables ou commercialiser les résultats, les connaissances et les compétences de la recherche* ». La valorisation des travaux de recherche est un enjeu important puisqu'elle conditionne en grande partie leur impact dans et en dehors du monde académique. Selon Collin-Lachaud and Michel (2020), quatre principales audiences peuvent être ciblées pour valoriser nos travaux : le milieu académique, les étudiants, les acteurs socio-économiques (privés et publics), et enfin les médias pour pouvoir toucher le grand public. Nous nous sommes principalement concentrés sur les deux premières, aussi la suite de cette section synthétise les actions que nous avons menées pour que nos recherches puissent être comprises et actionnables par ces audiences.

Le milieu académique

Au sein de la communauté académique, la valorisation prend essentiellement trois formes : la publication d'articles scientifiques (liste complète donnée en §A.5), la diffusion de ressources et d'outils (abordés plus haut dans les sections §5.1 et §5.2), et la participation à des conférences, ateliers et séminaires. C'est cette dernière forme de valorisation que nous détaillons ici.

Les participations aux campagnes d'évaluation sont un moyen efficace, que se soit en tant qu'organisateur pour amener d'autres chercheurs à s'intéresser à nos problématiques de recherche ou en tant que participant pour communiquer directement sur nos travaux. C'est ce que nous

avons fait avec l'organisation de l'atelier DÉfi Fouille de Texte (DEFT) 2016 sur la problématique de l'indexation de documents scientifiques en français (Daille et al., 2016) et la participation récurrente aux campagnes d'évaluation touchant à cette problématique (Boudin et al., 2012a; Bougouin et al., 2016c; Bouhandi et al., 2019).

Une autre action que nous avons entreprise pour accroître la visibilité de nos travaux sur la problématique de la production automatique de mots-clés a été l'organisation du tutoriel « *From Fundamentals to Recent Advances: A Tutorial on Keyphrasification* » à la conférence internationale ECIR 2022. Ce tutoriel, qui a permis de réunir plusieurs dizaines de personnes, est le résultat d'une collaboration avec deux collègues spécialistes du domaine : Rui Meng (Saleforce Research) et Debanjan Mahata (Moody's Analytics). À noter que l'ensemble des supports créés pour l'occasion (i. e. présentations, cahiers électroniques de démonstration (*notebooks*), enregistrements vidéo) ont été mis à disposition de la communauté.

Toujours dans l'optique de faire connaître nos travaux, nous avons animé plusieurs séminaires autour de la problématique de l'accès aux connaissances scientifiques. En particulier, nous avons présenté nos recherches lors de séminaires invités au *National Institute of Informatics* de Tokyo, et au ministère des armées pour les journées IA de défense.

Les étudiants

Les étudiants constituent une cible prioritaire auprès de qui valoriser nos recherches, il y va de notre capacité à recruter des candidats formés et motivés en thèse et en stage de Master. Parmi les actions que nous menons pour atteindre cet objectif, trois sortent du lot et s'inscrivent dans le cadre du parcours Apprentissage et Traitement Automatique des Langues (ATAL) du Master Informatique de Nantes Université : 1) la mise en place de cours proches de nos problématiques de recherche (e. g. méthode expérimentale, recherche d'information) ; 2) l'animation de séminaires réguliers de vulgarisation scientifique à destination des étudiants, les confrontant de manière informelle à nos pratiques de recherche ; et 3) l'accueil d'étudiants de M1 en stage d'initiation à la recherche au sein du laboratoire.

5.4. Discussion

Dans ce chapitre, nous avons synthétisé les efforts que nous avons engagé dans la construction de ressources langagières et d'outils logiciels, ainsi que les initiatives entreprises pour leur valorisation. Plusieurs défis restent à relever à court terme pour accélérer la recherche sur la production automatique de mots-clés. Tout d'abord, et nous l'avons déjà souligné en §5.1, il n'y a pas d'ensemble de données de grande taille dans une langue autre que l'anglais. Il existe pourtant des solutions pour certaines langues, comme par exemple pour le français où, dans le domaine des sciences humaines et sociales, des données pourraient être collectées à partir des bibliothèques numériques Persée ou OpenEdition. De plus, les

documents utilisés sont pour le moment cantonnés aux textes scientifiques et journalistiques. D'autres types de documents comme les emails ou les pages web semblent pourtant tout à fait pertinents et permettraient de mieux apprécier la valeur des modèles de production de mots-clés.

Ensuite, et malgré les nombreux travaux sur les méthodes de génération de mots-clés, peu de modèles entraînés sont véritablement disponibles pour la communauté scientifique. Davantage d'efforts sont nécessaires sur cet aspect d'autant plus que l'entraînement de ces modèles nécessite des quantités de calcul très importantes, de l'ordre de plusieurs jours de calcul avec des GPU (Ahmad et al., 2021). La plateforme Hugging Face, à travers son service de partage de modèles, est une solution technique intéressante pour résoudre ce problème mais peu de modèles de production de mots-clés y sont pour le moment disponibles.

For even the very wise cannot see all ends.

J.R.R. Tolkien, The Fellowship of the Ring

6

Conclusion et perspectives

Sommaire

Ce dernier chapitre clôture le manuscrit et donne quelques réflexions prospectives sur l'indexation par mots-clés et plus généralement sur les travaux de recherche émergeant de l'intersection des domaines du TAL et de la RI. Dans une première section (§6.1), nous poursuivons les discussions entamées dans les chapitres 2 et 3 et abordons quelques pistes de recherche sur l'indexation par mots-clés qui nous semblent importantes d'explorer à moyen terme. Dans une deuxième section (§6.2), nous discutons de la notion de sobriété numérique et décrivons plusieurs questions auxquelles nous pensons qu'il faudrait s'intéresser. Nous présentons dans une troisième section (§6.3) des travaux en cours sur l'aide à l'écriture scientifique et avançons quelques perspectives qui en découlent.

6.1. Indexation par mots-clés

L'éclosion en 2017 des méthodes neuronales a permis de franchir une nouvelle étape dans la qualité des mots-clés produits par rapport aux méthodes extractives classiques. Cependant, force est de constater que les résultats obtenus par les méthodes proposées depuis lors sont comparables et qu'un plateau de performance semble avoir été atteint (voir Figure 6.1). Il faut ajouter à cela l'inaptitude de ces méthodes à générer des mots-clés absents du texte source, point déjà soulevé en §3.4 et clairement illustré par la Figure 6.1 (courbe du bas). Surmonter cette limitation est un enjeu d'importance pour améliorer davantage les performances des méthodes de génération de mots-clés, d'autant plus que les mots-clés absents ont un effet prépondérant sur l'efficacité des systèmes de recherche d'information (Boudin and Gallina, 2021). Le récent glissement des méthodes de l'état-de-l'art vers des architectures de type *transformers*, notamment avec l'affinage (*fine-tuning*) de modèles

de langue génératifs pré-entraînés (Wu et al., 2022a,b), apporte un début de réponse à ce problème sans pour autant le résoudre.

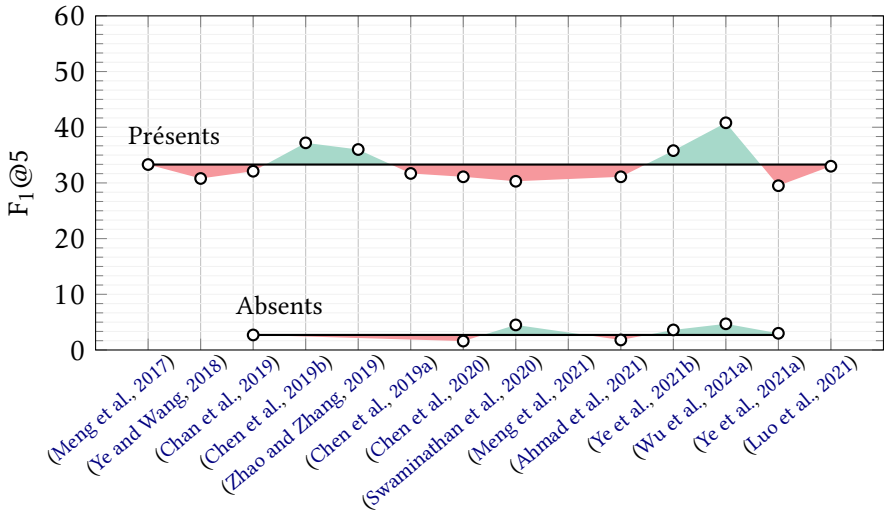


Figure 6.1. – Résultats en matière de $F_1@5$ des différents méthodes neuronales de génération de mots-clés proposées par la communauté sur l'ensemble de données KP20k. Les scores reportés sont directement repris des papiers.

Une partie de l'explication à cette situation de piétinement se trouve également au niveau des données. En effet, la mauvaise qualité des mots-clés de référence dans les ensembles de données disponibles pose la question de l'efficacité de l'entraînement des modèles. L'étude de méthodes non supervisées ou faiblement supervisées trouve ici tout son sens. Par exemple, l'utilisation des citations comme signal de supervision est une piste de recherche ancienne (Qazvinian et al., 2010; Caragea et al., 2014), mais qui redevient d'actualité avec la disponibilité de nouveaux ensembles de données de grande taille (Saier and Färber, 2020; Lo et al., 2020). Sur la même idée, une autre piste de recherche à explorer est celle de l'apprentissage actif en demandant, par exemple, aux utilisateurs de compléter et/ou de corriger les mots-clés associés aux résultats d'une requête. Une dernière piste de recherche, que nous explorons dans le projet ANR JCJC DELICES et qui fait le lien avec l'enjeu sur la génération des mots-clés absents, concerne les méthodes extractives non-supervisées et leur application au-delà du simple document (Wan and Xiao, 2008a; Boudin et al., 2020a), l'idée étant d'étendre l'extraction des mots-clés à un ensemble de documents similaires.

Au-delà des pistes de recherche évoquées jusqu'ici, il est intéressant de s'interroger sur la place des méthodes de génération de mots-clés à l'heure de la recherche d'information « neuronale ». Autrement dit, est-il toujours pertinent d'indexer des mots-clés générés alors que les systèmes de recherche d'information actuels utilisent des modèles de langue pré-entraînés pour l'appariement entre requêtes et documents ? Deux séries d'arguments sont en faveur d'une réponse positive :

1. la **latence** introduite par ces modèles est très élevée ; pour combattre ce problème, les systèmes sont construits sur des architectures multi-étages et utilisent les modèles neuronaux pour ré-ordonner les résultats d'une recherche initiale (Yates et al., 2021). Les méthodes de génération de mots-clés, et plus généralement d'expansion de documents, entraînent une amélioration significative des résultats de cette recherche initiale qui se cumule avec celle apportée par le ré-ordonancement (Nogueira et al., 2019; Thakur et al., 2021). De plus, elles s'appliquent en amont de l'indexation et ont un impact négligeable sur le temps de latence des requêtes.
2. l'**utilité** des mots-clés ne se limite pas à enrichir l'indexation ; les mots-clés constituent une interface d'accès naturelle, efficace et polyvalente au contenu des documents, et sont particulièrement pertinents dans un contexte d'exploration de données (Jones and Staveley, 1999; Chuang et al., 2012, inter alia). Ils ont l'avantage d'être intelligibles (et interprétables) et leur apport est bénéfique pour de nombreuses tâches de recherche d'information comme la recherche à facettes, la catégorisation de documents ou l'expansion de requêtes.

6.2. Sobriété numérique

Depuis l'émergence, au début des années 2010, de l'apprentissage profond nous constatons une augmentation rapide et substantielle de la taille des modèles utilisés en TAL et en RI. La Figure 6.2 illustre ce phénomène et montre que la taille des modèles (en nombre de paramètres) augmente d'un facteur dix chaque année. Ce constat va de pair avec la puissance de calcul nécessaire pour entraîner et utiliser ces modèles, qui elle aussi ne cesse de croître. Cette situation créé un problème environnemental majeur et met au ban une grande partie des chercheurs académiques qui ne peuvent avoir accès à de telles ressources calculatoires. Ainsi, la question de la sobriété numérique et plus concrètement de comment mener des travaux de recherche de qualité avec peu/moins de ressources est aujourd'hui prépondérante.

Une piste de recherche déjà bien balisée dans la communauté porte sur la distillation de connaissances (*knowledge distillation*) qui offre un moyen efficace de réduire la taille, et donc la latence et l'utilisation en mémoire/énergie des modèles (Hinton et al., 2015). Cependant, il ne s'agit que d'une réponse partielle à la question de la sobriété numérique puisqu'elle ne remet pas en cause l'entraînement du modèle initial et ajoute au contraire l'entraînement d'un second modèle au nombre de paramètres réduit. Une autre piste de travail concerne la simplification de l'architecture des modèles, et se place en contre-pied de la tendance actuelle qui est d'accroître la complexité des modèles pour améliorer les scores et prendre la tête des classements de performance de la communauté (e. g. GLUE, MS MARCO). Les efforts de recherche récents ont, pour la plupart, échoué à identifier les sources de gains empiriques dans les modèles, et par conséquent à justifier la complexité des modèles au-delà de l'amélioration des scores de référence (Moosavi et al., 2021).

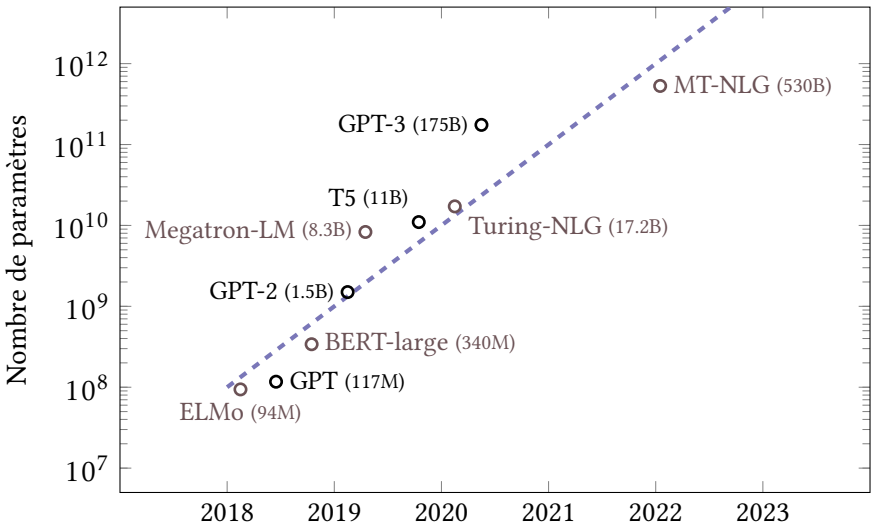


Figure 6.2. – Taille (en nombre de paramètres) des modèles de langue pré-entraînés.

Il faut tout de même noter que l'avenir n'est pas si sombre et que des solutions pratiques arrivent avec, du coté matériel, des unités de calcul dédiées plus puissantes et surtout moins énergivores (e. g. TPU v4 de Google, GPU A100 de NVIDIA) et, du coté logiciel, des bibliothèques mieux optimisées et une centralisation du stockage des modèles et des données qui tend à se généraliser (e. g. Hugging Face, Pytorch Hub). De même, dans les communautés TAL et RI, un effort considérable est engagé pour mobiliser les chercheurs autour des questions de sobriété numérique, notamment par l'organisation d'ateliers spécifiques (e. g. SustaiNLP, ReNeuIR) ou par l'ajout de thématiques dédiées au *Green and Sustainable NLP* dans les appels à communications des conférences.

6.3. Aide à l'écriture scientifique

Les travaux de recherche présentés dans ce manuscrit se concentrent sur l'amélioration de l'accès aux connaissances scientifiques sous le prisme de l'indexation par mots-clés. En parallèle et depuis 2019, nous nous intéressons à la problématique voisine de l'aide à l'écriture scientifique qui occupe aujourd'hui une part importante de notre activité. C'est donc tout naturellement que nous avançons ici quelques perspectives sur cette problématique, que nous articulons autour des deux axes présentés ci-dessous.

1. Les **expressions préétablies** (*formulaic expressions*) sont utilisées dans les articles scientifiques pour transmettre une fonction communicative, par exemple l'expression « *dans cet article, nous proposons* » a pour fonction de « *montrer l'objectif de l'article* » (Durrant and Mathews-Aydınlı, 2011). Dans le cadre d'une collaboration avec des collègues du National Institute of Informatics

(NII) à Tokyo, nous avons développé une méthode pour annoter automatiquement ces expressions ainsi que leurs fonctions (Iwatsuki et al., 2020, 2022). Comme l'usage des expressions préétablies diffère selon les domaines de recherche, une telle méthode permet la création rapide de ressources spécialisées pour la formulation de suggestions dans les systèmes d'aide à l'écriture (Iwatsuki and Aizawa, 2018). L'absence de cadre évaluatif de référence pour l'identification des expressions préétablies est sans doute le principal frein au développement de nouvelles méthodes et constitue aujourd'hui une perspective de recherche importante.

2. La **recommandation de citation** consiste à aider les chercheurs à choisir les articles à citer en proposant des citations appropriées pour un contexte donné (e. g. une phrase ou un paragraphe). Les collections de test disponibles pour cette tâche sont peu fiables car elles sont assemblées automatiquement à partir de textes extraits de fichiers PDF (Färber and Jatowt, 2020). Nous avons apporté une solution à ce problème par la création d'une collection de test dans laquelle les contextes et l'appariement des références ont été vérifiés manuellement (Boudin, 2021). Il reste cependant de nombreux défis à relever, notamment en ce qui concerne le caractère incomplet des jugements de pertinence, i. e. qui n'incluent pas les articles non cités mais pertinents. Une piste de recherche intéressante, qui emprunte au domaine de la scientométrie, serait d'explorer les réseaux de co-citations pour compléter les jugements de pertinence.

De manière plus large, l'analyse de la structure (argumentative, discursive) des articles scientifiques est un sujet de recherche particulièrement pertinent dans le cadre de l'aide à l'écriture scientifique. En effet, mettre en contraste la structure organisationnelle adoptée dans les textes d'un domaine de recherche avec celle d'un article en cours d'écriture permet de formuler des suggestions de réécriture personnalisées. Malgré de nombreux travaux sur le sujet (Teufel et al., 1999; Swales, 2004; Teufel et al., 2009, inter alia), plusieurs obstacles restent à lever pour permettre le développement de modèles performants et robustes, dont le premier est la rareté des données annotées. Cette problématique fait l'objet d'une thèse de doctorat qui a débutée fin 2022 et à laquelle je participe à l'encadrement.

Bibliographie

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004 : Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Wasi Ahmad, Xiao Bai, Soomin Lee, and Kai-Wei Chang. 2021. Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 1389–1404, Online. Association for Computational Linguistics.

Andrei Alexandrescu and Katrin Kirchhoff. 2007. Data-driven graph construction for semi-supervised graph-based learning in NLP. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, pages 204–211, Rochester, New York. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Bahman Bahmani, Abdur Chowdhury, and Ashish Goel. 2010. Fast incremental and personalized pagerank. *Proc. VLDB Endow.*, 4(3) :173–184.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1 : The 17th International Conference on Computational Linguistics*.

Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric SanJuan, and Xavier Tannier. 2016. Inex tweet contextualization task: Evaluation, results and lesson learned. *Information Processing & Management*, 52(5) :801–819.

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium. Association for Computational Linguistics.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Roi Blanco and Christina Lioma. 2007. Random walk term weighting for information retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 829–830, New York, NY, USA. Association for Computing Machinery.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3 :993–1022.

Stephen P. Borgatti and Martin G. Everett. 2006. A graph-theoretic perspective on centrality. *Social Networks*, 28(4) :466–484.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11) :2215–2222.

Florian Boudin. 2013a. A comparison of centrality measures for graph-based keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 834–838, Nagoya, Japan. Asian Federation of Natural Language Processing.

Florian Boudin. 2013b. TALN archives : a digital archive of French research articles in natural language processing (TALN archives : une archive numérique francophone des articles de recherche en traitement automatique de la langue) [in French]. In *Proceedings of TALN 2013 (Volume 2 : Short Papers)*, pages 507–514, Les Sables d'Olonne, France. ATALA.

Florian Boudin. 2015. Reducing over-generation errors for automatic keyphrase extraction using integer linear programming. In *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*, pages 19–24, Beijing, China. Association for Computational Linguistics.

Florian Boudin. 2016. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : System Demonstrations*, pages 69–73, Osaka, Japan. The COLING 2016 Organizing Committee.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 667–672, New Orleans, Louisiana. Association for Computational Linguistics.

Florian Boudin. 2021. Acm-cr : A manually annotated test collection for citation recommendation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2021*, JCDL '21, New York, NY, USA. Association for Computing Machinery.

Florian Boudin, Béatrice Daille, Évelyne Jacquey, and Jian-Yun Nie. 2020a. The DELICES project: Indexing scientific literature through semantic expansion. In *Proceedings of the First Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020*, volume 2621 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Florian Boudin and Ygor Gallina. 2021. Redefining absent keyphrases and their effect on retrieval effectiveness. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4185–4193, Online. Association for Computational Linguistics.

Florian Boudin, Ygor Gallina, and Akiko Aizawa. 2020b. Keyphrase generation for scientific document retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1126, Online. Association for Computational Linguistics.

Florian Boudin, Amir Hazem, Nicolas Hernandez, and Prajol Shrestha. 2012a. Participation du LINA à DEFT2012 (LINA at DEFT2012) [in French]. In *JEP-TALN-RECITAL 2012, Workshop DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge)*, pages 61–68, Grenoble, France. ATALA/AFCP.

Florian Boudin and Emmanuel Morin. 2013. Keyphrase extraction for n-best reranking in multi-sentence compression. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 298–305, Atlanta, Georgia. Association for Computational Linguistics.

Florian Boudin, Hugo Mougard, and Damien Cram. 2016. How document pre-processing affects keyphrase extraction performance. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 121–128, Osaka, Japan. The COLING 2016 Organizing Committee.

Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010a. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10(1) :1–6.

Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010b. Clinical information retrieval using document and PICO structure. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 822–830, Los Angeles, California. Association for Computational Linguistics.

Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010c. Deriving a test collection for clinical information retrieval from systematic reviews. In *Proceedings of the ACM Fourth International Workshop on Data and Text Mining in Biomedical Informatics*, DTMBIO '10, page 57–60, New York, NY, USA. Association for Computing Machinery.

Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010d. Positional language models for clinical information retrieval. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 108–115, Cambridge, MA. Association for Computational Linguistics.

Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2012b. Using a medical thesaurus to predict query difficulty. In *Advances in Information Retrieval*, pages 480–484, Berlin, Heidelberg. Springer Berlin Heidelberg.

Florian Boudin, Lixin Shi, and Jian-Yun Nie. 2010e. Improving medical information retrieval with pico element detection. In *Advances in Information Retrieval*, pages 50–61, Berlin, Heidelberg. Springer Berlin Heidelberg.

Adrien Bougouin. 2015. *Indexation automatique par termes-clés en domaines de spécialité*. Ph.D. thesis, Université de Nantes.

Adrien Bougouin, Sabine Barreaux, Laurent Romary, Florian Boudin, and Béatrice Daille. 2016a. TermITH-eval: a French standard-based resource for keyphrase extraction evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1924–1927, Portorož, Slovenia. European Language Resources Association (ELRA).

Adrien Bougouin and Florian Boudin. 2014. Topicrank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *Traitement Automatique des Langues*, 55(1) :45–69.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-based topic ranking for keyphrase extraction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 543–551, Nagoya, Japan. Asian Federation of Natural Language Processing.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2016b. Keyphrase annotation with graph co-ranking. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 2945–2955, Osaka, Japan. The COLING 2016 Organizing Committee.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2016c. Topicrank en domaines de spécialité : participation du lina à deft 2016. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. Volume 8 : DEFT*, pages 41–47, Paris, France. Association pour le Traitement Automatique des Langues. LINA at DEFT 2016.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2017a. Modélisation à base de graphe pour l'indexation en domaines de spécialité. *Recherche d'information, document et web sémantique*, 1(Numéro 1).

Adrien Bougouin, Florian Boudin, Béatrice Daille, Sabine Barreaux, Damien Cram, and Amir Hazem. 2017b. Indexation d'articles scientifiques présentation et résultats du défi fouille de textes deft 2016. *Recherche d'information, document et web sémantique*, 1(Numéro 1).

Mérième Bouhandi, Florian Boudin, and Ygor Gallina. 2019. DeFT 2019 : Auto-encodeurs, gradient boosting et combinaisons de modèles pour l'identification automatique de mots-clés. participation de l'équipe TALN du LS2N (autoencoders, gradient boosting and ensemble systems for automatic keyphrase assignment : The LS2N team participation's in the 2019 edition of DeFT). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Défi Fouille de Textes (atelier TALN-RECITAL)*, pages 57–66, Toulouse, France. ATALA.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1) :107–117. Proceedings of the Seventh International World Wide Web Conference.

Sam Brody, Sichao Wu, and Adrian Benton. 2021. Towards realistic few-shot relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5338–5345, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509 :257–289.

Erion Çano and Ondřej Bojar. 2019. Keyphrase generation: A text summarization struggle. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 666–672, Minneapolis, Minnesota. Association for Computational Linguistics.

Cornelia Caragea, Florin Adrian Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1435–1446, Doha, Qatar. Association for Computational Linguistics.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Hou Pong Chan, Wang Chen, Lu Wang, and Irwin King. 2019. Neural keyphrase generation via reinforcement learning with adaptive rewards. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2163–2174, Florence, Italy. Association for Computational Linguistics.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Jun Chen, Xiaoming Zhang, Yu Wu, Zhao Yan, and Zhoujun Li. 2018. Keyphrase generation with correlation constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4057–4066, Brussels, Belgium. Association for Computational Linguistics.

Wang Chen, Hou Pong Chan, Piji Li, Lidong Bing, and Irwin King. 2019a. An integrated approach for keyphrase generation via exploring the power of retrieval and extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2846–2856, Minneapolis, Minnesota. Association for Computational Linguistics.

Wang Chen, Hou Pong Chan, Piji Li, and Irwin King. 2020. Exclusive hierarchical decoding for deep keyphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1095–1105, Online. Association for Computational Linguistics.

Wang Chen, Yifan Gao, Jiani Zhang, Irwin King, and Michael R. Lyu. 2019b. Title-guided encoding for keyphrase generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01) :6268–6275.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. "without the clutter of unimportant words": Descriptive keyphrases for text visualization. *ACM Trans. Comput.-Hum. Interact.*, 19(3).

Grace Y Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC medical informatics and decision making*, 9(1) :1–13.

Vincent Claveau. 2020. *Du traitement des langues en recherche d'information et vice versa*. Habilitation à diriger des recherches, Univ. of Rennes.

Cyril Cleverdon. 1967. The cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Isabelle Collin-Lachaud and Géraldine Michel. 2020. Valoriser la recherche : une nouvelle mission des enseignants-chercheurs ? *Decisions Marketing*, 97(1) :5–16.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6) :407.

Andrew Collins and Joeran Beel. 2019. Document embeddings vs. keyphrases vs. terms for recommender systems: A large-scale online evaluation. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 130–133. IEEE Press.

Isaac Councill, C. Lee Giles, and Min-Yen Kan. 2008. ParsCit: an open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Béatrice Daille, Sabine Barreaux, Florian Boudin, Adrien Bougouin, Damien Cram, and Amir Hazem. 2016. Indexation d'articles scientifiques présentation et résultats du défi fouille de textes deft 2016. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016. Volume 8 : DEFT*, pages 1–12, Paris, France. Association pour le Traitement Automatique des Langues. Automatic indexing of scientific papers.

Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. 2007. The identification of clinically important elements within medical journal abstracts: Patient_population_problem, exposure_intervention, comparison, outcome, duration and results (pecodr). *Journal of Innovation in Health Informatics*, 15(1) :9–16.

Dina Demner-Fushman, Barbara Few, Susan E. Hauser, and George Thoma. 2006. Automatically Identifying Health Outcome Information in MEDLINE Records. *Journal of the American Medical Informatics Association*, 13(1) :52–60.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1) :63–103.

Romain Deveaud and Florian Boudin. 2012. Lia/lina at the inex 2012 tweet contextualization track. In *INitiative for the Evaluation of XML Retrieval (INEX)*.

Romain Deveaud and Florian Boudin. 2013a. Contextualisation automatique de tweets à partir de wikipédia. In *Conférence en Recherche d'Information et Applications (CORIA)*.

Romain Deveaud and Florian Boudin. 2013b. Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization. In *INitiative for the Evaluation of XML Retrieval (INEX)*.

Romain Deveaud and Florian Boudin. 2014. De quoi parle ce tweet? résumer wikipédia pour contextualiser des microblogs. *The Information - Intelligence - Interaction (I3) Journal*, pages 37–56.

Romain Deveaud, Florian Boudin, and Patrice Bellot. 2011a. Lia at inex 2010 book track. In *Comparative Evaluation of Focused Retrieval*, pages 118–127, Berlin, Heidelberg. Springer Berlin Heidelberg.

Romain Deveaud, Florian Boudin, Eric SanJuan, and Patrice Bellot. 2011b. Correction de césures et enrichissement de requêtes pour la recherche de livres. In *Conférence en Recherche d'Information et Applications (CORIA)*, pages 89–96.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shizhe Diao, Ruijia Xu, Hongjin Su, Yilei Jiang, Yan Song, and Tong Zhang. 2021. Taming pre-trained language models with n-gram representations for low-resource domain adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 3336–3349, Online. Association for Computational Linguistics.

Zhuoye Ding, Qi Zhang, and Xuanjing Huang. 2011. Keyphrase extraction from online news using binary integer programming. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 165–173, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Duy Dinh, Lynda Tamine, and Fatiha Boubekeur. 2013. Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. *Artificial Intelligence in Medicine*, 57(2) :155–167.

Philip Durrant and Julie Mathews-Aydınlı. 2011. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1) :58–72.

Miles Efron, Peter Organisciak, and Katrina Fenlon. 2012. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 911–920, New York, NY, USA. Association for Computing Machinery.

Samhaa R. El-Beltagy and Ahmed Rafea. 2010. KP-miner: Participation in SemEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 190–193, Uppsala, Sweden. Association for Computational Linguistics.

Güneş Erkan and Dragomir R. Radev. 2004. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371, Barcelona, Spain. Association for Computational Linguistics.

Leonhard Euler. 1741. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

J. Fagan. 1987. Automatic phrase indexing for document retrieval. In *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '87, page 91–101, New York, NY, USA. Association for Computing Machinery.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *Int. J. Digit. Libr.*, 21(4) :375–405.

Zichu Fei, Qi Zhang, and Yaqian Zhou. 2021. Iterative GNN-based decoder for question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Felice Ferrara, Nirmala Pudota, and Carlo Tasso. 2011. A keyphrase-based paper recommender system. In *Digital Libraries and Archives*, pages 14–25, Berlin, Heidelberg. Springer Berlin Heidelberg.

Katja Filippova. 2010. Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 322–330, Beijing, China. Coling 2010 Organizing Committee.

Corina Florescu and Cornelia Caragea. 2017. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11) :964–971.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Ygor Gallina, Florian Boudin, and Béatrice Daille. 2020. Large-scale evaluation of keyphrase extraction models. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 271–278, New York, NY, USA. Association for Computing Machinery.

Bin Gao, Tie-Yan Liu, Wei Wei, Taifeng Wang, and Hang Li. 2011. Semi-supervised ranking on very large graphs with rich metadata. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 96–104, New York, NY, USA. Association for Computing Machinery.

Jianfeng Gao, Patrick Nguyen, Xiaolong(Shiao-Long) Li, Chris Thrasher, Mu Li, and Kuansan Wang. 2010. A comparative study of bing web n-gram language models for web search and natural language processing. In *Proceeding of the 33rd Annual ACM SIGIR Conference*. Association for Computing Machinery, Inc.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Andrea Glaser and Hinrich Schütze. 2012. Automatic generation of short informative sentiment summaries. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 276–285, Avignon, France. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Wentian Guo, Yuchen Li, Mo Sha, and Kian-Lee Tan. 2017. Parallel personalized pagerank on dynamic graphs. *Proc. VLDB Endow.*, 11(1) :93–106.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael Gusenbauer and Neal R. Haddaway. 2020. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods*, 11(2) :181–217.

Carl Gutwin, Gordon Paynter, Ian Witten, Craig Nevill-Manning, and Eibe Frank. 1999. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems*, 27(1) :81–104.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1) :10–18.

Khaled M. Hammouda, Diego N. Matute, and Mohamed S. Kamel. 2005. Corephrase : Keyphrase extraction for document clustering. In *Machine Learning and Data Mining in Pattern Recognition*, pages 265–274, Berlin, Heidelberg. Springer Berlin Heidelberg.

Juhyun Han, Taehwan Kim, and Joongmin Choi. 2007. Web document clustering by using automatic keyphrase extraction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, WI-IATW '07, page 56–59, USA. IEEE Computer Society.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1262–1273, Baltimore, Maryland. Association for Computational Linguistics.

Taher H. Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, page 517–526, New York, NY, USA. Association for Computing Machinery.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. *LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation*, page 639–648. Association for Computing Machinery, New York, NY, USA.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the*

Association for Computational Linguistics, pages 712–721, Gothenburg, Sweden. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Maël Houbre, Florian Boudin, and Béatrice Daille. 2022. A large-scale dataset for biomedical keyphrase generation. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis*, Online. Association for Computational Linguistics.

Eduard Hovy and Chin-Yew Lin. 1998. Automated text summarization and the Summarist system. In *TIPSTER TEXT PROGRAM PHASE III : Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 197–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Chien-yu Huang, Arlene Casey, Dorota Głowacka, and Alan Medlar. 2019. Holes in the outline: Subject-dependent abstract quality and its implications for scientific literature search. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 289–293, New York, NY, USA. Association for Computing Machinery.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 53–64, Vancouver, British Columbia. Association for Computational Linguistics.

Yongjie Huang and Meng Yang. 2021. Breadth first reasoning graph for multi-hop question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 5810–5821, Online. Association for Computational Linguistics.

Anette Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 216–223.

Anette Hulth and Beáta B. Megyesi. 2006. A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 537–544, Sydney, Australia. Association for Computational Linguistics.

Kenichi Iwatsuki and Akiko Aizawa. 2018. Using formulaic expressions in writing assistance systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2678–2689, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. 2020. An evaluation dataset for identifying communicative functions of sentences in English scholarly papers. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1712–1720, Marseille, France. European Language Resources Association.

Kenichi Iwatsuki, Florian Boudin, and Akiko Aizawa. 2022. Extraction and evaluation of formulaic expressions used in scholarly papers. *Expert Systems with Applications*, 187 :115840.

Aishwarya Jadhav and Vaibhav Rajan. 2018. Extractive summarization with SWAP-NET: Sentences and words from alternating pointer networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 142–151, Melbourne, Australia. Association for Computational Linguistics.

Karen Sparck Jones. 1999. *What is the Role of NLP in Text Retrieval?*, pages 1–24. Springer Netherlands, Dordrecht.

Steve Jones and Gordon W Paynter. 2003. An evaluation of document keyphrase sets. *Journal of Digital Information*, 4(1).

Steve Jones and Mark S. Staveley. 1999. Phrasier: A system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 160–167, New York, NY, USA. Association for Computing Machinery.

Noriko Kando. 2001. Overview of the second ntcir workshop. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*.

Jin Young Kim and W. Bruce Croft. 2012. A field relevance model for structured document retrieval. In *Advances in Information Retrieval*, pages 97–108, Berlin, Heidelberg. Springer Berlin Heidelberg.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2009. Extracting domain-specific words - a statistical approach. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 94–98, Sydney, Australia.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. SemEval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26, Uppsala, Sweden. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction. Technical report, University of Trento.

Bruce Krulwich and Chad Burkey. 1996. Learning user information interests through extraction of semantically significant phrases. In *Proceedings of the AAAI spring symposium on machine learning in information access*, pages 110–112.

Esther Landhuis. 2016. Scientific literature : Information overload. *Nature*, 535(7612) :457–458.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Joongbo Shin, and Kyomin Jung. 2021. KPQA: A metric for generative question answering using keyphrase weights. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2105–2115, Online. Association for Computational Linguistics.

Chi-Hong Leung and Wing-Kay Kan. 1997. A statistical learning approach to automatic indexing of controlled index terms. *Journal of the American Society for Information Science*, 48(1) :55–66.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jimmy Lin. 2009. Is searching full text more effective than searching abstracts ? *BMC bioinformatics*, 10(1) :1–15.

Jimmy Lin. 2021. The neural hype, justified! a recantation. *SIGIR Forum*, 53(2) :88–93.

Jimmy Lin. 2022. A proposed conceptual framework for a representational approach to information retrieval. *SIGIR Forum*, 55(2).

Aldo Lipani. 2016. Fairness in information retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1171, New York, NY, USA. Association for Computing Machinery.

Aldo Lipani. 2019. On biases in information retrieval models and evaluation. *SIGIR Forum*, 52(2) :172–173.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Coling 2008 : Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization*, pages 17–24, Manchester, UK. Coling 2008 Organizing Committee.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2021. Highlight-transformer: Leveraging key phrase aware attention to improve abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pages 5021–5027, Online. Association for Computational Linguistics.

Zhiyuan Liu, Xinxiong Chen, Yabin Zheng, and Maosong Sun. 2011. Automatic keyphrase extraction by bridging vocabulary gap. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 135–144, Portland, Oregon, USA. Association for Computational Linguistics.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376, Cambridge, MA. Association for Computational Linguistics.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 257–266, Singapore. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Patrice Lopez and Laurent Romary. 2010. HUMB: Automatic key term extraction from scientific articles in GROBID. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 248–251, Uppsala, Sweden. Association for Computational Linguistics.

Yichao Luo, Yige Xu, Jiacheng Ye, Xipeng Qiu, and Qi Zhang. 2021. Keyphrase generation with fine-grained evaluation-guided reinforcement learning. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, pages 497–507, Punta Cana, Dominican Republic. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 8(3) :243–281.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Daniel Marcu. 1997. The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chap-*

ter of the Association for Computational Linguistics, pages 96–103, Madrid, Spain. Association for Computational Linguistics.

Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and João P. Neto. 2012. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 399–403, Istanbul, Turkey. European Language Resources Association (ELRA).

Hiroshi Maruyama. 1990. Structural disambiguation with constraint propagation. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 31–38, Pittsburgh, Pennsylvania, USA. Association for Computational Linguistics.

Olena Medelyan and Ian H. Witten. 2006. Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '06, page 296–297, New York, NY, USA. Association for Computing Machinery.

Olena Medelyan and Ian H. Witten. 2008. Domain-independent automatic keyphrase indexing with small training sets. *Journal of the American Society for Information Science and Technology*, 59(7) :1026–1040.

Rui Meng, Debanjan Mahata, and Florian Boudin. 2022. From fundamentals to recent advances : A tutorial on keyphrasification. In *Advances in Information Retrieval*, pages 582–588, Cham. Springer International Publishing.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4985–5007, Online. Association for Computational Linguistics.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 582–592, Vancouver, Canada. Association for Computational Linguistics.

Donald Metzler and W Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 170–173, Barcelona, Spain. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. PageRank on semantic networks, with application to word sense disambiguation. In *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, pages 1126–1132, Geneva, Switzerland. COLING.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language : Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

R. V. Mises and H. Pollaczek-Geiringer. 1929. Praktische verfahren der gleichungsauflösung . *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 9(2) :152–164.

Ioannis Mitliagkas, Michael Borokhovich, Alexandros G. Dimakis, and Constantine Caramanis. 2015. Frogwild! fast pagerank approximations on graph engines. *Proc. VLDB Endow.*, 8(8) :874–885.

Nafise Sadat Moosavi, Iryna Gurevych, Angela Fan, Thomas Wolf, Yufang Hou, Ana Marasović, and Sujith Ravi, editors. 2021. *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Virtual.

Aurélie Névéol, Kelly Zeng, and Olivier Bodenreider. 2006. Besides precision & recall : Exploring alternative approaches to evaluating an automatic indexing tool for medline. In *AMIA Annual Symposium Proceedings*, volume 2006, page 589. American Medical Informatics Association.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pages 317–326, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS 2017 Workshop Autodiff*.

Krutarth Patel and Cornelia Caragea. 2021. Exploiting position and contextual word embeddings for keyphrase extraction from scientific papers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 1585–1591, Online. Association for Computational Linguistics.

Fuchun Peng and Andrew McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics : HLT-NAACL 2004*, pages 329–336, Boston, Massachusetts, USA. Association for Computational Linguistics.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 275–281, New York, NY, USA. Association for Computing Machinery.

Marco Ponza, Luciano Del Corro, and Gerhard Weikum. 2018. Facts that matter. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1043–1048, Brussels, Belgium. Association for Computational Linguistics.

Animesh Prasad and Min-Yen Kan. 2019. Glocal: Incorporating global information in local convolution for keyphrase extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1837–1846, Minneapolis, Minnesota. Association for Computational Linguistics.

Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. 2010. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 895–903, Beijing, China. Coling 2010 Organizing Committee.

Delip Rao and David Yarowsky. 2009. Ranking and semi-supervised classification on large scale graphs using map-reduce. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 58–65, Suntec, Singapore. Association for Computational Linguistics.

Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and Jimmy Lin. 2019. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5370–5381, Hong Kong, China. Association for Computational Linguistics.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question : a key to evidence-based decisions. *Acp j club*, 123(3) :A12–A13.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 42–49, New York, NY, USA. Association for Computing Machinery.

Stephen E Robertson, Steve Walker, Micheline Beaulieu, and Peter Willett. 1999. Okapi at trec-7 : automatic ad hoc, filtering, vlc and interactive track. *Nist Special Publication SP*, pages 253–264.

David L. Sackett. 1997. Evidence-based medicine. *Seminars in Perinatology*, 21(1) :3–5. Fatal and Neonatal Hematology for the 21st Century.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023) :71–72.

Tarek Saier and Michael Färber. 2020. Unarxive: A large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics*, 125(3) :3085–3108.

T.y.s.s Santosh, Debarshi Kumar Sanyal, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. SaSAKE: Syntax and semantics aware keyphrase extraction from research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5372–5383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1) :61–80.

Connie Schardt, Martha B Adams, Thomas Owens, Sheri Keitz, and Paul Fontelo. 2007. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC medical informatics and decision making*, 7(1) :1–6.

Natalie Schluter. 2014. Centrality measures for non-contextual graph-based unsupervised single document keyword extraction. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, pages 455–460, Marseille, France. Association pour le Traitement Automatique des Langues.

Alexander Thorsten Schutz. 2008. Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods. *Master's thesis, National University of Ireland*.

Federico Scozzafava, Marco Maru, Fabrizio Brignone, Giovanni Torrisi, and Roberto Navigli. 2020. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 37–46, Online. Association for Computational Linguistics.

Xianjie Shen, Yinghan Wang, Rui Meng, and Jingbo Shang. 2021. Unsupervised deep keyphrase generation.

Jieming Shi, Renchi Yang, Tianyuan Jin, Xiaokui Xiao, and Yin Yang. 2019. Realtime top-k personalized pagerank over large graphs on gpus. *Proc. VLDB Endow.*, 13(1) :15–28.

Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. OCR++: A robust framework for information extraction from scholarly articles. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, pages 3390–3400, Osaka, Japan. The COLING 2016 Organizing Committee.

Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104 :333–339.

Luciana B Sollaci and Mauricio G Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure : a fifty-year survey. *Journal of the medical library association*, 92(3) :364.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4) :521–544.

Amy M Steier and Richard K Belew. 1993. Exporting phrases : A statistical analysis of topical language. In *Second Symposium on Document Analysis and Information Retrieval*, pages 179–190.

Lucas Sterckx, Thomas Demeester, Johannes Deleu, and Chris Develder. 2015. Topical word importance for fast keyphrase extraction. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 121–122, New York, NY, USA. Association for Computing Machinery.

C Rockelle Strader. 2009. Author-assigned keywords versus library of congress subject headings. *Library resources & technical services*, 53(4) :243–250.

Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri : A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligent analysis*, pages 2–6.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 78–88, Melbourne, Australia. Association for Computational Linguistics.

Zhiqing Sun, Jian Tang, Pan Du, Zhi-Hong Deng, and Jian-Yun Nie. 2019. Divgraphpointer: A graph pointer network for extracting diverse keyphrases. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 755–764, New York, NY, USA. Association for Computing Machinery.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

John M. Swales. 2004. *Research Genres: Explorations and Applications*. Cambridge Applied Linguistics. Cambridge University Press.

Avinash Swaminathan, Haimin Zhang, Debanjan Mahata, Rakesh Gosangi, Rajiv Ratn Shah, and Amanda Stent. 2020. A preliminary exploration of GANs for keyphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8021–8030, Online. Association for Computational Linguistics.

Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 407–414, New York City, USA. Association for Computational Linguistics.

Nedelina Teneva and Weiwei Cheng. 2017. Salience rank: Efficient keyphrase extraction with topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 530–535, Vancouver, Canada. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Simone Teufel et al. 1999. *Argumentative zoning : Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh Edinburgh, Scotland.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Antoine Tixier, Fragkiskos Malliaros, and Michalis Vazirgiannis. 2016. A graph degeneracy-based approach to keyword extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1860–1870, Austin, Texas. Association for Computational Linguistics.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 33–40, Sapporo, Japan. Association for Computational Linguistics.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Peter D. Turney. 1999. Learning to extract keyphrases from text. *NRC Technical Report ERB-l 057. National Research Council, Canada*, pages 1–43.

Ana Sabina Uban, Cornelia Caragea, and Liviu P. Dinu. 2021. Studying the evolution of scientific topics and their relationships. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pages 1908–1922, Online. Association for Computational Linguistics.

Nicola Ueffing, Franz Josef Och, and Hermann Ney. 2002. Generation of word graphs in statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 156–163. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xiaojun Wan and Jianguo Xiao. 2008a. CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 969–976, Manchester, UK. Coling 2008 Organizing Committee.

Xiaojun Wan and Jianguo Xiao. 2008b. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 855–860. AAAI Press.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic. Association for Computational Linguistics.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4) :641–649.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. Kea: Practical automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, page 254–255, New York, NY, USA. Association for Computing Machinery.

Di Wu, Wasi Uddin Ahmad, Sunipa Dev, and Kai-Wei Chang. 2022a. Representation learning for resource-constrained keyphrase generation.

Huanqin Wu, Wei Liu, Lei Li, Dan Nie, Tao Chen, Feng Zhang, and Di Wang. 2021a. UniKeyphrase: A unified extraction and generation framework for keyphrase prediction. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pages 825–835, Online. Association for Computational Linguistics.

Huanqin Wu, Baijiaxin Ma, Wei Liu, Tao Chen, and Dan Nie. 2022b. Fast and constrained absent keyphrase generation by prompt-based learning. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021b. *Self-Supervised Graph Learning for Recommendation*, page 726–735. Association for Computing Machinery, New York, NY, USA.

Jianxin Yang, Wenge Rong, Libin Shi, and Zhang Xiong. 2019a. Sequential Attention with Keyword Mask Model for Community-based Question Answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2201–2211, Minneapolis, Minnesota. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019b. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1129–1132, New York, NY, USA. Association for Computing Machinery.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 44th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, SIGIR '21, page 2666–2668, New York, NY, USA. Association for Computing Machinery.

Hai Ye and Lu Wang. 2018. Semi-supervised learning for neural keyphrase generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4142–4153, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Ye, Ruijian Cai, Tao Gui, and Qi Zhang. 2021a. Heterogeneous graph neural networks for keyphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2705–2715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiacheng Ye, Tao Gui, Yichao Luo, Yige Xu, and Qi Zhang. 2021b. One2Set: Generating diverse keyphrases as a set. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 4598–4608, Online. Association for Computational Linguistics.

Chenhan Yuan and Hoda Eldardiry. 2021. Unsupervised relation extraction: A variational autoencoder approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1929–1938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2020. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975, Online. Association for Computational Linguistics.

Hongyuan Zha. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 113–120, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai. 1997. Fast statistical parsing of noun phrases for document indexing. In *Fifth Conference on Applied Natural Language Processing*, pages 312–319, Washington, DC, USA. Association for Computational Linguistics.

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 403–410, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2) :179–214.

Yuan Zhang, Dong Wang, and Yan Zhang. 2019. Neural ir meets graph embedding: A ranking model for product search. In *The World Wide Web Conference*, WWW '19, page 2390–2400, New York, NY, USA. Association for Computing Machinery.

Jing Zhao and Yuxiang Zhang. 2019. Incorporating linguistic constraints into keyphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5224–5233, Florence, Italy. Association for Computational Linguistics.

Fanwei Zhu, Yuan Fang, Kevin Chen-Chuan Chang, and Jing Ying. 2013. Incremental and accuracy-aware personalized pagerank through scheduled approximation. *Proc. VLDB Endow.*, 6(6) :481–492.

A

Curriculum Vitæ

A.1. Formation et expérience professionnelle

Formation, titres et diplômes

2004 **Licence Génie Mathématique et Informatique**
Université d'Avignon / University of Birmingham, UK (ERASMUS)

2006 **Master Recherche Informatique**
Université d'Avignon
Parcours Traitement Automatique du Langage Naturel Ecrit et Oral

2008 **Doctorat Informatique**
Université d'Avignon
Titre : Exploration d'approches statistiques pour le résumé automatique de texte
Sous la direction de Juan-Manuel Torres-Moreno et Marc El-Bèze

Parcours professionnel et activités d'enseignement

2006→08 **Assistant de recherche**
Université d'Avignon / Laboratoire Informatique d'Avignon
Enseignements à l'IUT d'Avignon

2009→10 **Chercheur post-doctoral**
Université de Montréal / Recherche Appliquée en Linguistique Informatique
Enseignements à l'Université de Montréal

2010→11 **Attaché temporaire d'enseignement et de recherche**
Université d'Avignon / Laboratoire Informatique d'Avignon

Depuis 2011 **Maître de conférences**

Enseignement : rattaché au département d'informatique de Nantes Université

Recherche : rattaché au Laboratoire d'Informatique Nantes Atlantique (LINA, UMR 6241) de 2011 à 2016, au Laboratoire des Sciences du Numérique à Nantes (LS2N, UMR 6004) depuis janvier 2017

2018→21 Prime d'Encadrement Doctoral et de Recherche (PEDR)

juil.→déc. 2019 Séjour de recherche au National Institute of Informatics (NII), Tokyo

A.2. Encadrements

La Table A.1 donne une vue synthétique de mes activités d'encadrement. On notera l'encadrement de 4 thèses (2 soutenues) et de 26 étudiants de Master et Licence.

Niveau	D	M2	M1	L3
Nombre étudiants	4	4	20	2

Table A.1. – Nombre d'étudiants encadrés (ou en cours d'encadrement) par niveau.

Encadrements de thèse

1. **Adrien Bougouin** (*début : 01 octobre 2012, soutenance : 27 octobre 2015*)
   Co-encadrement avec Béatrice Daille à hauteur de 50%
   Titre : Indexation automatique par termes-clés en domaines de spécialité.
   Publications : 10 publications dont trois articles de revues (Bougouin and Boudin, 2014; Bougouin et al., 2017b,a), une conférence internationale de rang CORE A (Bougouin et al., 2016b) et une de rang CORE B (Bougouin et al., 2013).
   Devenir : Ingénieur logiciel chez Wovn Technologies, Tokyo

2. **Ygor Gallina** (*début : 01 octobre 2018, soutenance : 28 mars 2022*)
   Co-encadrement avec Béatrice Daille à hauteur de 50%
   Titre : Indexation de bout-en-bout dans les bibliothèques numériques scientifiques.
   Publications : 6 publications dont deux conférences internationales de rang CORE A* (Gallina et al., 2020; Boudin et al., 2020b), une de rang CORE A (Boudin and Gallina, 2021) et une de rang CORE B (Gallina et al., 2019).
   Devenir : Attaché temporaire d'enseignement et de recherche, Nantes Université

3. **Maël Houbre** (*début : 25 mars 2022*)
   Co-encadrement avec Béatrice Daille à hauteur de 50%
   Titre : Génération non-supervisée de mots-clés absents pour l'indexation d'articles scientifiques

4. **Léane Jourdan** (*début : 01 octobre 2022*)
   Co-encadrement avec Nicolas Hernandez et Richard Dufour à hauteur de 30%.
   Titre : Neural approaches for modelling the argumentative structure of Research articles

Encadrements de stages de M2

1. **Rémi Bois** (2014)
   Sujet : Multi-document summarization through sentence fusion.

2. **Carol Couillerot** (2019)
   Sujet : État de l'art de la détection de mails d'hameçonnage par apprentissage automatique.

3. **Timothée Poulain** (2020)
   Sujet : Generating absent keyphrases for scientific document indexing.
4. **Rima Boubeker** (2022, co-encadrée avec Richard Dufour à hauteur de 50%)
   Sujet : Generation automatique de hashtags pour des messages courts issus de Twitter.

Encadrements de stages de M1

1. **Hugo Mougard**, **Grégoire Jadi**, **Rémi Bois** et **Noémi Salaün** (2013, co-encadré avec Fabien Poulard à hauteur de 50%)
   Sujet : Semi-supervised extraction of content from the Web.
2. **Robin Boncorps**, **Guillaume Charon**, **Gwendal Daniel** et **Jérôme Pagès** (2013, co-encadré avec Emmanuel Morin à hauteur de 50%)
   Sujet : Compression multi-phrases en contexte multilingue.
3. **Loïc Jankowiak** (2013)
   Sujet : Étude des paramètres d'entrée pour la compression multi-phrase.
4. **Adeline Granet** et **Alexis Linard** (2014)
   Sujet : Détection automatique des tweets humoristiques.
5. **Marie Lenogue**, **Anthony Pena** et **Clément Tek** (2015, co-encadré avec Hugo Mougard à hauteur de 50%)
   Sujet : Quels concepts pour le résumé automatique par extraction ?
6. **Naïma Mazri** et **Mathis Yassin** (2020, co-encadré avec Emmanuel Morin à hauteur de 50%)
   Sujet : Créer une collection de test pour la recherche d'information.
7. **Ronan Belleil** et **Sylvain Beaudoin** (2021)
   Sujet : Recommandation de citation : étendre les références aux textes plein.
8. **Antoine Jamelot** (2022)
   Sujet : Entraîner un modèle neuronal de génération de texte avec peu de données et de calculs.
9. **Leïla Brehon** (2022)
   Sujet : Améliorer la compréhension de texte pour les personnes dyslexiques avec des mots-clés.

Encadrements de stages de L3

1. **Rémi Bois** (2012, co-encadré avec Nicolas Hernandez à hauteur de 50%)
   Sujet : Intégration d'un bibliothèque de mesures de similarités textuelles au sein du framework UIMA.
2. **Thomas Fonteneau** (2013)
   Sujet : Compression de phrases par transduction d'arbres.

A.3. Responsabilités scientifiques

Animation équipes de recherche

Depuis 2021 **Responsable adjoint de l'équipe Traitement Automatique du Langage Naturel (TALN)** du Laboratoire des Sciences du Numérique de Nantes, actuellement composée de 11 permanents (4 PR, 7 MCF) et de 8 doctorants.

Contrats de recherche

2012→16 **Membre partenaire du projet ANR TermITH** (Terminologie et Indexation de Textes en sciences Humaines) porté par Evelyne Jacquey (ATILF, UMR 7118) et en partenariat avec l'INIST, le LIDILEM (EA 609) et l'INRIA.

2015 **Porteur du projet CNRS-PEPS INS2I/INSMI GOLEM** (Approche par Optimisation pour l'Extraction de Mots-clés) en partenariat avec l'équipe SLP de l'IRCCyN (UMR 6597).
- 6 000€ pour l'achat de matériel.

2016 **Porteur du projet CNRS-PEPS INS2I TALIAS** (Le TAL au service de l'indexation des articles scientifiques).
- 5 000€ pour l'achat de matériel.

2018 **Porteur du projet collaboratif avec l'entreprise Dantooin Technologies** via CAPACITÉS, filiale de valorisation de la recherche de Nantes Université.
- 2 656€ pour le financement de stagiaires.

2019 **Porteur du projet AtlanSTIC 2020 IKEBANA** (Improving Keyphrase Extraction By Adopting Neural Architectures) en collaboration avec le NII de Tokyo.
- 20 000€ pour un séjour de recherche (6 mois).

2020→25 **Porteur du projet ANR JCJC DELICES** (Indexer la littérature scientifique par expansion sémantique) en collaboration avec Béatrice Daille (LS2N - Nantes Université), Evelyne Jacquey (CNRS, Université de Lorraine) et Jian-Yun Nie (RALI, Université de Montréal).
- 193 968 €, englobant le financement d'un doctorant et d'un post-doctorant (1 an).

2020 **Porteur du projet AtlanSTIC 2020 WASP** (Building a Writing Assistance system for Scientific Papers) en collaboration avec le NII de Tokyo.
- 10 000€ pour la visite d'un chercheur étranger (6 mois, annulé en raison de la COVID-19).

2023→26 **Porteur du projet CNRS-DGA-AID NaviTerm** (Navigation terminologique pour une montée en compétence rapide et personnalisée sur un domaine de recherche).
- 245 125€, englobant le financement d'un doctorant et d'un ingénieur de recherche (6 mois).

A.4. Animation de la recherche

Expertises

Évaluation de projet ANR pour l'Appel à Projet Générique (AAPG)

⊢ 2016, 2017, 2020, 2021

Revue à mi-parcours de projets ANR

⊢ 2019 (CE23 - Données, Connaissances, Big data, Contenus multimédias, Intelligence Artificielle)

Activités éditoriales

La liste suivante énumère les comités auxquels j'ai participé depuis 2010. J'ai été *area chair* de la conférence majeure du domaine du TAL en 2021 (ACL, rang CORE A*) et je suis *action editor* (équivalent à *area chair*) pour le ACL Rolling Review (système centralisé de relecture de papiers de la communauté internationale du TAL) depuis octobre 2021. Notamment, j'ai été nommé parmi les meilleurs relecteurs de la conférence internationale NAACL 2016 (rang CORE A) et désigné relecteur exceptionnel (*outstanding reviewer*) pour la conférence internationale EMNLP 2020 (rang CORE A).

→ *Area chairing* de conférences internationales

Annual Meeting of the Association for Computational Linguistics (ACL) (2021), Conference on Empirical Methods in Natural Language Processing (EMNLP) (2022), Action Editor pour le ACL Rolling Review (depuis oct. 2021)

→ Comités de lecture de conférences internationales

Annual Meeting of the Association for Computational Linguistics (ACL) (2017, 2019, 2020, 2021), Conference on Empirical Methods in Natural Language Processing (EMNLP) (2019, 2020, 2021), Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2016, 2018, 2019, 2021), Conference of the European Chapter of the Association for Computational Linguistics (EACL) (2021), European Conference on Information Retrieval (ECIR) (2022), International Conference on Computational Linguistics (COLING) (2012, 2016, 2018, 2020, 2022), International Joint Conference on Natural Language Processing (IJCNLP) (2017), Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL) (2020, 2022), International Conference on Language Resources and Evaluation (LREC) (2018, 2020, 2022), Conference on Information and Knowledge Management (CIKM) (2011)

→ Comités de lecture de conférences nationales

Conférence Traitement Automatique des Langues Naturelles (TALN) (2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017), Rencontre des Étudiants Chercheurs en Informatique pour le TAL (RECITAL) (2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017), COnférence en Recherche d'Information et Applications (CORIA) (2017)

→ Comités de lecture d'ateliers internationaux

Automatic Summarization for Creative Writing (Creative-Summ) (2022 @ COLING), MultiLing Workshops (Summarization and summary evaluation across source types and genres) (2017 @ EACL, 2019 @ RANLP, 2020 @ COLING), New Frontiers in Summarization (NewSum) (2019 @ EMNLP, 2021 @ EMNLP), International Workshop on Narrative Extraction from Texts (Text2Story) (2019 @ ECIR, 2021 @ ECIR, 2022 @ ECIR), Digital Infrastructures for Scholarly Content Objects

(DISCO) (2021 @ JCDL), Artificial Intelligence for Narratives Workshop (AI4Narratives) (2020 @ IJCAI-PRICAI) Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE) (2020 @ ACL), Workshop on Building and Using Comparable Corpora (BUCC) (2015 @ LREC)

→ Comités de lecture de revues

Artificial Intelligence (AI) en 2021 ; Information Processing & Management (IPM) en 2015, 2018 (2×) ; WIREs Data Mining and Knowledge Discovery (WIDM) en 2018 ; Natural Language Engineering (NLE) en 2018 ; Cognitive Systems Research (CSR) en 2018 ; Recherche d'Information, Document et Web Sémantique (RIDoWS) en 2017 ; Language Resources and Evaluation (LRE) en 2017 ; Information Retrieval (IR) en 2014, Traitement Automatique des Langues (TAL) en 2014, 2016, 2018 ; Journal of Natural Language Engineering (JNLE) en 2016 ; Journal of Cheminformatics (J Cheminform) en 2016

Participation jurys de thèse

2018 Examinateur de thèse de Elvys Linhares Pontes (LIA, Université d'Avignon)

2018 Examinateur de thèse de Marco Basaldella (AI Laboratory, University of Udine)

2018 Examinateur de thèse de Rashedur Rahman (LIMSI, Université Paris Sud)

2021 Examinateur de thèse de Guokan Shang (DaSciM, Institut polytechnique de Paris)

Responsabilités et activités au sein des sociétés savantes ou associations

2013→16 **Membre élu du conseil d'administration de l'ATALA** (Association pour le Traitement Automatique des Langues). Créateur de TALN Archives, une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue (http://talnarchives.atala.org/).

2018 **Participation au pré-GDR TAL** - axe productions langagières. Contributions à la rédaction du pré-rapport sur les grands enjeux du TAL en vue de la création du GDR.

2019→Auj. **Membre du comité de pilotage du collège TLH** (Technologies du Langage Humain) de l'AFIA (Association Française pour l'Intelligence Artificielle) dont la mission consiste, entre autres, à soutenir l'organisation de manifestations scientifiques et de communiquer autour des recherches des communautés françaises du TAL, de la RI et de l'IA (≈ 2 réunions / an).

Organisation colloques, conférences, journées d'étude

2013 **Président du comité d'organisation** et programme des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), orga-

nisées conjointement à la conférence nationale Traitement Automatique des Langues Naturel (TALN) (≈ 200 participants).

2022 **Membre du comité d'organisation** de la journée commune AFIA-THL / Association de Recherche d'Information et Applications (ARIA) / GDR TAL sur le thème de l'accès interactif à l'information (≈ 50 participants).

2022 **Membre du comité d'organisation** du tutoriel *From Fundamentals to Recent Advances : A Tutorial on Keyphrasification* à la conférence *European Conference on Information Retrieval* (ECIR).

2020→Auj. **Responsable du Petit Séminaire ATAL** pour le parcours ATAL (Apprentissage et Traitement Automatique de la Langue) du Master Informatique de Nantes Université (≈ 20 participants, 6 séminaires / an).

A.5. Publications

La Table A.2 donne une vue synthétique de ma production scientifique. On notera 15 publications dans des conférences internationales majeures (rang CORE A*/A). Selon Google Scholar (consulté le 09/09/2022), mes indicateurs bibliométiques sont : 1 607 citations, indice h : 18 et indice i-10 : 30. Mes 5 articles les plus cités sont [24] (354 citations), [22] (156 citations), [2] (139 citations), [16] (138 citations) et [23] (88 citations).

Type de production	**Nombre**
Revues internationales à comité de lecture	2
Revues nationales à comité de lecture	4
Conférences internationales avec comité de lecture	28*
Ateliers internationaux avec comité de lecture	11
Conférences nationales avec comité de lecture	9
Ateliers nationaux avec comité de lecture	4
Chapitres d'ouvrage	1
Manuscrit de thèse de doctorat	1
Total	60

Table A.2. – Liste classée par type du nombre d'articles publiés. * la distribution des articles de conférences internationales par rang CORE est A* : 3, A : 12, B : 7 et C : 6.

La suite de cette section dresse la liste des publications auxquelles j'ai participé. Mon nom est mentionné **en gras**, et ceux des étudiants que j'ai encadrés sont soulignés.

Articles dans revues internationales à comité de lecture

[1] Kenichi Iwatsuki, **Florian Boudin** et Akiko Aizawa. Extraction and evaluation of formulaic expressions used in scholarly papers. *Expert Systems with Applications*, 187. 2022.

[2] **Florian Boudin**, Jian-Yun Nie, Joan Bartlett, Roland Grad, Pierre Pluye et Martin Dawes. Combining classifiers for robust PICO element detection. *BMC Medical Informatics and Decision Making*, 10 :29. 2010.

Articles dans revues nationales à comité de lecture

[3] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. Modélisation à base de graphe pour l'indexation en domaines de spécialité. *Recherche d'information, document et web sémantique*, 17 :1. 2017.

[4] Adrien Bougouin, **Florian Boudin**, Béatrice Daille, Sabine Barreaux, Damien Cram et Amir Hazem. Indexation d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT 2016. *Recherche d'information, document et web sémantique*, 17 :1. 2017.

[5] Adrien Bougouin et **Florian Boudin**. TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *Traitement Automatique des Langues*, 55 :1. 2014.

[6] Romain Deveaud et **Florian Boudin**. De quoi parle ce Tweet ? Résumer Wikipédia pour contextualiser des microblogs. *Information-Intelligence-Interaction Journal*. 2014.

Articles de conférences internationales avec comité de lecture

[7] Amir Hazem, Mérieme Bouhandi, **Florian Boudin** et Béatrice Daille. Cross-lingual and Cross-domain Transfer Learning for Automatic Term Extraction from Low Resource Data. *Language Resources and Evaluation Conference (LREC)*. 2022.

[8] Rui Meng, Debajan Mahata et **Florian Boudin**. From Fundamentals to Recent Advances : A Tutorial on Keyphrasification. *European Conference on Information Retrieval (ECIR)*. 2022.

[9] **Florian Boudin**. ACM-CR : A Manually Annotated Test Collection for Citation Recommendation. *Joint Conference on Digital Libraries (JCDL)*. 2021.

[10] **Florian Boudin** et Ygor Gallina. Redefining Absent Keyphrases and their Effect on Retrieval Effectiveness. *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*. 2021.

[11] **Florian Boudin**, Béatrice Daille, Evelyne Jacquey et Jian-Yun Nie. The DELICES Project : Indexing Scientific Literature Through Semantic Expansion. *Joint Conference of the Information Retrieval Communities in Europe (CIRCLE)*. 2020.

[12] **Florian Boudin**, Ygor Gallina et Akiko Aizawa. Keyphrase Generation for Scientific Document Retrieval. *Association for Computational Linguistics (ACL)*. 2020.

[13] Ygor Gallina, **Florian Boudin** et Béatrice Daille. Large-Scale Evaluation of Keyphrase Extraction Models. *Joint Conference on Digital Libraries (JCDL)*. 2020.

[14] Kenichi Iwatsuki, **Florian Boudin** et Akiko Aizawa. An Evaluation Dataset for Identifying Communicative Functions of Sentences in English Scholarly Papers. *Language Resources and Evaluation Conference (LREC)*. 2020.

[15] Ygor Gallina, **Florian Boudin** et Béatrice Daille. KPTimes : A Large-Scale Dataset for Keyphrase Generation on News Documents. *International Conference on Natural Language Generation (INLG)*. 2019.

[16] **Florian Boudin**. Unsupervised Keyphrase Extraction with Multipartite Graphs. *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*. 2018.

[17] **Florian Boudin**. pke : an open source python-based keyphrase extraction toolkit. *International Conference on Computational Linguistics (COLING)*. 2016.

[18] Adrien Bougouin, Sabine Barreaux, Laurent Romary, **Florian Boudin** et Béatrice Daille. TermITH-Eval : a French Standard-Based Resource for Keyphrase Extraction Evaluation. *Language Resources and Evaluation Conference (LREC)*. 2016.

[19] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. Keyphrase Annotation with Graph Co-Ranking. *International Conference on Computational Linguistics (COLING)*. 2016.

[20] **Florian Boudin**, Hugo Mougard et Benoit Favre. Concept-based Summarization using Integer Linear Programming : From Concept Pruning to Multiple Optimal Solutions. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2015.

[21] Ophélie Lacroix, Denis Béchet et **Florian Boudin**. Label Pre-annotation for Building Non-projective Dependency Treebanks for French. *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2014.

[22] **Florian Boudin**. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction. *International Joint Conference on Natural Language Processing (IJCNLP)*. 2013.

[23] **Florian Boudin** et Emmanuel Morin. Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression. *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*. 2013.

[24] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction. *International Joint Conference on Natural Language Processing (IJCNLP)*. 2013.

[25] **Florian Boudin**, Jian-Yun Nie et Martin Dawes. Using a Medical Thesaurus to Predict Query Difficulty. *European Conference on Information Retrieval (ECIR)*. 2012.

[26] **Florian Boudin**, Stéphane Huet et Juan-Manuel Torres-Moreno. A Graph-based Approach to Cross-language Multi-document Summarization. *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2011.

[27] **Florian Boudin**, Jian-Yun Nie et Martin Dawes. Clinical Information Retrieval using Document and PICO Structure. *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*. 2010.

[28] **Florian Boudin**, Jian-Yun Nie et Martin Dawes. Positional Language Models for Clinical Information Retrieval. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2010.

[29] **Florian Boudin**, Lixin Shi et Jian-Yun Nie. Improving Medical Information Retrieval with PICO Element Detection. *European Conference on Information Retrieval (ECIR)*. 2010.

[30] **Florian Boudin**, Marc El-Bèze et Juan-Manuel Torres-Moreno. A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization. *International Conference on Computational Linguistics (COLING)*. 2008.

[31] **Florian Boudin**, Juan Torres-Moreno et Marc El-Bèze. Mixing Statistical and Symbolic Approaches for Chemical Names Recognition. *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2008.

[32] **Florian Boudin**, Juan-Manuel Torres-Moreno et Patricia Velázquez-Morales. An Efficient Statistical Approach for Automatic Organic Chemistry Summarization. *International Conference on Natural Language Processing (GoTAL)*. 2008.

[33] **Florian Boudin** et Juan Torres Moreno. NEO-CORTEX : A Performant User-Oriented Multi-Document Summarization System. *Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. 2007.

[34] **Florian Boudin** et Juan-Manuel Torres-Moreno. A Cosine Maximization Minimization approach for User Oriented Multi-Document Update Summarization. *Recent Advances in Natural Language Processing (RANLP)*. 2007.

Articles d'ateliers internationaux avec comité de lecture

[35] Maël Houbre, **Florian Boudin** et Béatrice Daille. A large-scale dataset for biomedical keyphrase generation. 13th International Workshop on Health Text Mining and Information Analysis (LOUHI). 2022.

[36] Amir Hazem, Mérieme Bouhandi, **Florian Boudin** et Béatrice Daille. TermEval 2020 : TALN-LS2N System for Automatic Term Extraction. *6th International Workshop on Computational Terminology (CompuTerm)*. 2020.

[37] **Florian Boudin**, Hugo Mougard et Damien Cram. How Document Pre-processing affects Keyphrase Extraction Performance. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. 2016.

[38] **Florian Boudin**. Reducing Over-generation Errors for Automatic Keyphrase Extraction using Integer Linear Programming. *Workshop on Novel Computational Approaches to Keyphrase Extraction*. 2015.

[39] Emmanuel Morin, Amir Hazem, **Florian Boudin** et Elizaveta Loginova-Clouet. LINA : Identifying Comparable Documents from Wikipedia. *Eighth Workshop on Building and Using Comparable Corpora*. 2015.

[40] Romain Deveaud et **Florian Boudin**. Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization. *INitiative for the Evaluation of XML Retrieval (INEX)*. 2013.

[41] Romain Deveaud et **Florian Boudin**. LIA/LINA at the INEX 2012 Tweet Contextualization track. *INitiative for the Evaluation of XML Retrieval (INEX)*. 2012.

[42] Romain Deveaud, **Florian Boudin** et Patrice Bellot. LIA at INEX 2010 Book Track. *INitiative for the Evaluation of XML Retrieval (INEX)*. 2011.

[43] **Florian Boudin**, Jian-Yun Nie et Martin Dawes. Deriving a test collection for clinical information retrieval from systematic reviews. *Data and Text Mining in Biomedical Informatics (DTMBIO)*. 2010.

[44] **Florian Boudin**, Marc El-Bèze et Juan-Manuel Torres-Moreno. The LIA Update Summarization system at TAC-2008. *Text Analysis Conference (TAC)*. 2008.

[45] **Florian Boudin**, Benoit Favre, Frederic Béchet, Marc El-Bèze, Laurent Gillard et Juan-Manuel Torres-Moreno. The LIA-Thales summarization system at DUC-2007. *Document Understanding Conference (DUC)*. 2007.

[46] Benoit Favre, Frederic Béchet, Patrice Bellot, **Florian Boudin**, Marc El-Beze, Laurent Gillard, Guy Lapalme et Juan-Manuel Torres-Moreno. The LIA-Thales summarization system at DUC-2006. *Document Understanding Conference (DUC)*. 2006.

Articles de conférences nationales avec comité de lecture

[47] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. Modélisation unifiée du document et de son domaine pour une indexation par termes-clés libre et contrôlée. *Traitement Automatique des Langues Naturelles (TALN)*. 2016.

[48] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. Influence des domaines de spécialité dans l'extraction de termes-clés. *Traitement Automatique des Langues Naturelles (TALN)*. 2014.

[49] **Florian Boudin**. TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue. *Traitement Automatique des Langues Naturelles (TALN)*. 2013.

[50] Romain Deveaud et **Florian Boudin**. Contextualisation automatique de Tweets à partir de Wikipédia. *Conférence en Recherche d'Information et Applications (CORIA)*. 2013.

[51] Nicolas Hernandez et **Florian Boudin**. Construction d'un large corpus écrit libre annoté morpho-syntaxiquement en français. *Traitement Automatique des Langues Naturelles (TALN)*. 2013.

[52] **Florian Boudin** et Nicolas Hernandez. Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank. *Traitement Automatique des Langues Naturelles (TALN)*. 2012.

[53] Romain Deveaud, **Florian Boudin**, Eric SanJuan et Patrice Bellot. Correction de césures et enrichissement de requêtes pour la recherche de livres. *Conférence en Recherche d'Information et Applications (CORIA)*. 2011.

[54] Stéphane Huet, **Florian Boudin** et Juan-Manuel Torres-Moreno. Utilisation d'un score de qualité de traduction pour le résumé multi-document cross-lingue. *Traitement Automatique des Langues Naturelles (TALN)*. 2011.

[55] **Florian Boudin** et Juan-Manuel Torres-Moreno. Résumé automatique multi-document et indépendance de la langue : une première évaluation en français. *Traitement Automatique des Langues Naturelles (TALN)*. 2009.

Articles d'ateliers nationaux avec comité de lecture

[56] Mérième Bouhandi, **Florian Boudin** et Ygor Gallina. DeFT 2019 : Auto-encodeurs, Gradient Boosting et combinaisons de modèles pour l'identification automatique de mots-clés. *Défi Fouille de Textes (DEFT)*. 2019.

[57] Adrien Bougouin, **Florian Boudin** et Béatrice Daille. TopicRank en domaines de spécialité : participation du LINA à DEFT 2016. *Défi Fouille de Textes (DEFT)*. 2016.

[58] Béatrice Daille, Sabine Barreaux, **Florian Boudin**, Adrien Bougouin, Damien Cram et Amir Hazem. Indexation d'articles scientifiques : Présentation et résultats du défi fouille de textes DEFT 2016. *Défi Fouille de Textes (DEFT)*. 2016.

[59] **Florian Boudin**, Amir Hazem, Nicolas Hernandez et Prajol Shrestha. Participation du LINA à DEFT 2012. *Défi Fouille de Textes (DEFT)*. 2012.

Chapitres d'ouvrages

[60] **Florian Boudin** et Juan-Manuel Torres-Moreno. A Maximization-Minimization Approach for Update Text Summarization. *Current Issues in Linguistic Theory : Recent Advances in Natural Language Processing*. 2009.

Autres

[61] **Florian Boudin**. Exploration d'approches statistiques pour le résumé automatique de texte. *Thèse de doctorat*. Laboratoire Informatique d'Avignon – Université d'Avignon, 2008

Nantes Université

LS2N

Titre : Analyse et indexation de textes scientifiques

Mots clés : recherche d'information, traitement automatique des langues, indexation automatique par mots-clés, textes scientifiques, méthodes de graphes, évaluation, aide à l'écriture scientifique

Résumé : Les travaux présentés dans cette habilitation à diriger des recherches (HDR) ont pour objet l'analyse et l'indexation des textes scientifiques, et se situent à la croisée de deux thématiques de recherche : celle du Traitement Automatique des Langues (TAL) qui concerne l'analyse, la compréhension et la production de langage naturel, et celle de la Recherche d'Information (RI) qui étudie la manière de retrouver des informations dans une collection de documents. Nous nous intéressons à la problématique de la recherche bibliographique, c'est-à-dire la recherche de documents dans la littérature scientifique (e.g. articles, ouvrages, thèses) en rapport avec un sujet d'étude, et plus particulièrement à l'enrichissement des métadonnées associées aux documents pour en améliorer l'accessibilité et la diffusion.

Nos travaux concernent le développement de méthodes automatisées de génération de mots-clés dont la singularité réside dans l'utilisation de méthodes de graphes et d'algorithmes d'ordonnancement de sommets. Nous nous penchons sur la problématique de l'évaluation indirecte des mots-clés générés au travers de tâches applicatives et de leur exploitation dans les moteurs de recherche et de recommandation académique. Nous présentons les travaux que nous avons menés dans la construction de ressources langagières, le développement d'outils logiciels et leur valorisation dans la communauté scientifique. Nous terminons par quelques réflexions prospectives sur l'indexation par mots-clés et plus généralement sur les travaux de recherche émergeant de l'intersection des thématiques du TAL et de la RI.

Title: Analysing and indexing scientific texts

Keywords: information retrieval, natural language processing, keyword indexing, scientific texts, graph-based methods, evaluation, scientific writing assistance

Abstract: The work presented in this "Habilitation à Diriger des Recherches" (Accreditation to Supervise Research) focuses on the analysis and indexing of scientific texts and lies at the intersection of two research themes: Natural Language Processing (NLP), which involves the analysis, understanding, and generation of natural language, and Information Retrieval (IR), which studies ways to retrieve information from a collection of documents. We are interested in the question of scholarly document retrieval, which involves searching for documents in the scientific literature (e.g., articles, books, theses) related to a specific subject of study. More specifically, our research aims to enhance the metadata associated with documents to improve their accessibility and dissemination.

Our work focuses on the development of automated methods for keyword generation, which are characterized by the unique utilization of graph-based techniques and node ranking algorithms. We delve into the issue of indirectly evaluating automatically generated keywords through application-specific tasks and their utilization in search engines and academic recommendation systems. We present our efforts into constructing linguistic resources, developing software tools, and their dissemination within the scientific community. Finally, we conclude with some prospective insights into keyword indexing and, more broadly, the emerging research at the intersection of NLP and IR themes.